COMPUTERIZED ADAPTIVE TESTING:
A CASE STUDY

Robert Samuel Kayler

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

COMPUTERIZED ADAPTIVE TESTING:

A CASE STUDY

by

Robert Samuel Kayler

December 1980

Thesis Advisor: R. A. Weitzman

Approved for public release; distribution unlimited.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Computerized Adaptive Testing: A Case Study		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis; December 1980 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert Samuel Kayler		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE December 1980 13. NUMBER OF PAGES 122
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Naval Postgraduate School Monterey, California 93940		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Intelligence Testing
Mental Testing
AFQT
Armed Force Qualification Testing [See page 2]

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This thesis is a case study of mental testing in the military as it applies to mental qualifications for service. The thesis begins with a review of the literature concerning the history of mental testing, particularly in the military services, through the current Armed Services Vocational Aptitude Battery (ASVAB) forms -8, -9, and -10. Then, a discussion of issues facing mental testing in general is

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

BLOCK 19. KEY WORDS (Continued)

ASVAB
Armed Services Vocational Aptitude Battery
Adaptive Testing
Computerized Adaptive Testing
Military Testing

BLOCK 20. ABSTRACT (Continued)

presented, followed by a report of research into Computerized Adaptive Testing (CAT) currently conducted at the Navy Personnel Research and Development Center, San Diego. Finally, a concluding chapter discusses some considerations involved in the implementation of CAT.

Approved for public release; distribution unlimited

Computerized Adaptive Testing: A Case Study

by

Robert Samuel Kayler
Lieutenant Commander, Medical Service Corps,
United States Navy
B.B.A., George Washington University, 1966

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
December 1980

ABSTRACT

This thesis is a case study of mental testing in the military as it applies to mental qualifications for service. The thesis begins with a review of the literature concerning the history of mental testing, particularly in the military services, through the current Armed Services Vocational Aptitude Battery (ASVAB) forms -8, -9, and -10. Then, a discussion of issues facing mental testing in general is presented, followed by a report of research into Computerized Adaptive Testing (CAT) currently conducted at the Navy Personnel Research and Development Center, San Diego. Finally, a concluding chapter discusses some considerations involved in the implementation of CAT.

TABLE OF CONTENTS

I. INTRODUCTION

Since World War I, American educators and institutional leaders have applied a number of techniques to the business of selecting and placing personnel in training and education programs and in jobs. Some of the methods employed in seeking the best choice from among several applicants for school or work include the use of references, interviews, past performance records, practical tests, probationary periods, and apprenticing. The most well known means of assessing abilities today, however, is the psychological test.

The reason for all of the "screening" and attempting to assess the abilities of individuals is that there are a number of differences. People differ in intelligence, skills, abilities to learn, motivation, and experience. When an employer hires a worker or when a college admits a student, there is an expectation that the worker will possess the ability to do at least a certain minimum standard of work; the student is expected to "pass" his or her school work. The skill and ability requirements for jobs and educational programs vary and there is competition among aspirants for the "best" jobs or for admission to the "most prestigious" colleges and universities so that it becomes necessary to make decisions about who will do what jobs or attend which schools.

In a utopian world, our goal would be to place all workers into positions ideally suited for them, to admit students to the precise education and training which both meets their motivational needs and fulfills a need for society. We strive in our complex society to make the best choices around whom to bring together, in which institutions, to make what happen, in whose interpretation of what is best for the individual, the institution, or society. The best selection assumes that individuals so chosen will apply their abilities, skills, and motivations in the best possible way for all concerned. It also assumes that our institutions will make the best use of our available human resources.

We are finding the task of best selection and placement difficult because of the differences among people. As if it were not a complicated enough problem dealing with estimations of individual differences, we find that there are a myriad interpretations of what the differences are, how they should be assessed, and who should have the power to assess them. In the midst of the turmoil over individual likeness and differences -- societal debate concerning racial discrimination, bias, sexism, etc. -- we find the psychological testing industry.

A complete discussion of the complex problems, techniques, and developments in the psychological testing industry is beyond the scope of intent in this thesis. What follows is

a case-study approach to psychological testing in the military services. As will be seen, the military sector has played a significant role in the development of techniques that attempt to measure differences among people. We specifically want to focus on an innovative approach that is currently being researched and developed.

The thesis begins with two chapters on the history of mental testing. In Chapter II, we trace the history of testing through World War I. Chapter III describes military entrance-qualification testing through the current tests. Chapter IV discusses some of the issues which face the testing industry. In Chapter V, we describe the research which is being conducted into computerized adaptive testing (CAT). Finally, a brief concluding chapter offers some considerations for implementation of computerized testing.

II. HISTORY OF MENTAL TESTING THROUGH 1917

Modern psychometrics has its roots in the ancient Chinese practice of rigorous examinations leading to public service. The full range of psychological testing as we know it has grown from three earlier developments: civil service examinations; the assessment of academic achievement in schools, colleges and universities; and studies of individual differences by Western scientists.

A. CIVIL SERVICE TESTING

Modern psychological measurement has its roots in the ancient Chinese culture where, for more than 3,000 years, an elaborate system of competitive examinations was used to select personnel for government positions in China. Origins of the system go back to 2200 B.C., when the Chinese emperor examined his officials every three years to determine their fitness for continuing in office. The significance of the Chinese contribution has prompted one psychologist/author to dedicate his research "To those wise men of China who, thousands of years ago, invented the psychological test" [Dubois 1970].

The Chinese system of examinations for public service was modified down through the years. In 1115 B.C., candidates for government positions were examined for their proficiency in music, archery, horsemanship, writing, arithmetic, and the rites and ceremonies of public and private life. In the Han

Dynasty (202 B.C. - 200 A.D.), written examinations tested knowledge of civil law, military affairs, agriculture, taxation, and the geography of the Empire.

The examination system in China took its final form about 1370 A.D., at which time proficiency in remembering and interpreting the Confucian classics was emphasized. The examinations were rigorous, the final series being given in the capital city, Peking. Only three percent of the final applicants were successful in becoming Mandarins, eligible for public office. This was a great distinction, for where thousands presented themselves for examination, only hundreds passed [Brubacher 1947].

European contacts with China in the sixteenth century led to the introduction of a system of examinations in France in 1791. However, Napoleon abolished the system and it remained out of service for many years thereafter.

In 1883, the United Kingdom first used competitive examinations to aid in the selection of trainees for civil service in India. Interest in the use of competitive examinations for civil service spread from this British beginning to America in the 1860's. Dorman B. Eaton (1823-1899) was active in civil service reform in the United States and became president of the Civil Service Board under President Grant. Eaton became a permanent member of the Civil Service Commission established under the Civil Service Act of January 16, 1883 [Dubois, 1970].

B. UNIVERSITY AND SCHOOL TESTING

Ancient Greek and Roman schools are not known to have used formal examinations.[1] Also, from the standpoint of positive contributions to a scientific psychology, the Middle Ages are relatively unimportant. Cathedral and monastery schools of the era used no formal examinations. For hundreds of years, university exams were exclusively oral. Medieval teaching methods were based upon the formal lecture which would be memorized by students. Written examinations were first introduced at the University of Bologna in 1219. Examinations were strict; when a candidate was examined for the Doctor's degree, the process frequently lasted for a week or more [Mayer 1933]. Louvain University introduced competitive examinations as early as 1441 in which students were divided into four classes: "rigorosi" (honor men), "transibiles" (satisfactory), "gratiosi" (charity passes), and failures [Dubois 1970].

Development of written examinations followed the introduction of paper. Members of the Jesuit Order, founded in

---

[1]Anaxagoras (c. 500 - c. 428 B.C.) is credited with introducing the notion that the order of the world must be explained as well as its constituents, and the ordering principles he found in nous, something corresponding to human intelligence or reason, but not yet contrasted to matter. (Nous is synonymous with mind or reason.) Anaxagoras' philosophy is significant in that it points out the psychological processes for special attention [Heidbreder 1933].

1540 by St. Ignatius of Loyola, were pioneers in the systematic use of written tests, both for the placement of students and for their evaluation after instruction [McGucken 1932]. Written examinations were introduced at Oxford and Cambridge Universities around 1800. Printed question papers were introduced in 1828. The University of London was chartered in 1836 to examine candidates for degrees from two London colleges, and it later opened its examinations to externs as well [Dubois 1970: 10].

By the mid-1800's, written examinations had been widely recognized for their usefulness as a basis for important decisions such as who should be permitted to exercise a profession, who should be awarded degrees, and who should serve in public positions.

Thus the stage was set for the beginnings of scientific psychology as it pertains to the testing and measuring of mental abilities. Steps had been taken toward the development of uniformity of testing and objectivity of assessment. What followed was the extension of measurement to other areas of human behavior.

C. SCIENTIFIC STUDY OF HUMAN DIFFERENCES

1. English Contributions

Sir Francis Galton (1822-1911) became a principal founder of the scientific study of human differences. He engaged in a variety of studies of individual differences

including psychology, photography and human faculty with word associations [Forrest 1974].

Galton, who was originally trained in medicine at Kings College, London, Trinity College, Cambridge, and at hospitals in Birmingham and London, abandoned his medical studies for a period of time after inheriting a comfortable fortune at age 22. Later his enthusiasm for intellectual pursuits led him to more than half a century of wide-ranging creative scholarship [Dubois 1970]. Stimulated by discussions with Croom-Robertson, the first editor of the periodical, Mind, Galton proceeded at his own expense to equip and open an anthropometric laboratory as part of the International Health Exhibition in South Kensington in 1884 [Forrest 1974].

Visitors to Galton's anthropometric laboratory could pay a three-penny admission fee and have a series of tests and measurements recorded. The visitors were furnished copies of the results which were filed on record for reference. Measurements included standing height, sitting height, arm span, weight, vital breathing capacity, strength of pull, strength of squeeze, swiftness of blow, keenness of sight, memory of form, discriminations of odor and steadiness of hand. Nearly 10,000 people passed through this laboratory before it was closed and moved to the South Kensington Museum in 1885.

The significance of Galton's work in the anthropometric laboratory lies in its contribution to statistical

methods. Galton utilized some of the data in developing tables of percentile norms, by sex, for several physical and behavioral characteristics, including height, weight, strength, and keenness of sight. The most important outcome was his development of the concepts of regression and correlation as a tool for understanding imperfect relationships between variables [Heidbreder 1933].

Karl Pearson (1857-1936), a student of Galton's, followed the work of his teacher, further developed theories of statistics, and worked out the now famous and widely used product-moment formula for obtaining the coefficient of correlation [Heidbreder 1933]. Pearson also developed techniques for deriving multiple correlation, worked out methods for finding correlations from four-fold tables, biserial r, and the chi-square test for goodness of fit. Charles Spearman (1863-1945), an English psychologist who improved upon Pearson's work and made significant contributions to psychometrics, will be discussed below.

2. Introduction of Testing in America

James McKeen Cattell (1860-1944) introduced the Galton tradition of testing to the United States. An American psychologist, Cattell used Galton's methods at the University of Pennsylvania in 1888 and at Columbia University in 1891. Cattell's tests were largely of sensory and motor functions with related measures of perception, association and memory beginning to appear. He utilized simple apparatus, scoring in physical units of measure such as time, distance, pitch,

temperature and force. Most of the measures were obviously related to the experimental psychology of the day, which emphasized the study of sensation, reaction time, and discrimination [Cattell 1947].

Cattell was one of the founders of the American Psychological Association (APA) and of early psychological journals. He was the world's first Professor of Psychology and was influential in the development of notable psychologists such as Lightner Witmer (1867-1956), E. L. Thorndike (1874-1949), R. S. Woodworth (1869-1962), F. L. Wells (1884-1964), and E. K. Strong, Jr. (1884-1963).

Lightner Witmer was the founder of the first psychological clinic in America at the University of Pennsylvania in 1896. Edward L. Thorndike authored the first book in psychological statistics, An Introduction to the Theory of Mental and Social Measurements, in 1904. Robert S. Woodworth went on to pioneer in the study of race differences and became the author of the first personality inventory in 1918. E. K. Strong, Jr., developed the Strong Vocational Interest Blank in 1927 [Dubois 1970].

3. German Contributions

Cattell also studied for a period of three years in Leipzig under Wilhelm Wundt in the world's first psychological laboratory,[2] founded by Wundt in 1879 [Heidbreder 1933].

[2]William James' laboratory, established at Harvard in 1875, did not take on significant psychological research characteristics until some years later.

While little in the measurement of individual differences developed directly from this first laboratory, several contributions evolved from Wundt's pupils.

Emil Kraepelin (1855-1926), a psychiatrist who had been one of Wundt's first pupils, inaugurated comparative psychological testing of the sane and insane. Kraepelin and his associates proposed a comprehensive system of comparative testing of the sane and insane that would consider individual characteristics such as mental ability, trainability, memory, sensitivity, fatigability, ability to recover from fatigue, depth of sleep, and distractibility. He recognized the need for standardization of testing procedures and the necessity to repeat examinations of each case a sufficient number of times so that chance variations could be excluded [White 1964].

Axel Oehrn, Adolf Gross, and Joseph Reis, students of Kraepelin, sought to obtain normative data on healthy individuals, with whom the mentally ill could be compared. The experience in testing in which Kraepelin and his associates engaged undoubtedly gave them insights into mental abnormality, but results were often disappointing [White 1964; Spearman 1904].

Herman Ebbinghaus (1850-1909) achieved a major breakthrough in testing techniques; he invented the completion test. As early as 1897, in a study of the possible effects of fatigue and of the most satisfactory arrangement of working hours for school children, Ebbinghaus used the

completion test, which has since proved to be one of the most useful testing techniques [Heidbreder 1933].

Ebbinghaus' completion test consisted of passages of texts with words and/or parts of words omitted and with each omission indicated by a line. The student's task was to "complete" as many of the missing parts as possible in a limited time span. Ebbinghaus pointed out the ease with which completion tests could be scored and compared among individuals to obtain a numerical assessment of their respective intellectual abilities [Dubois 1970].

While several psychologists used the completion technique in their investigations, including Lewis M. Terman [Terman 1916] and Binet (discussed below), the practical importance of the group method of administering a psychological test was not recognized until later.

D. DEVELOPMENT OF INDIVIDUAL SCALES

It has been argued that ability measurement began with the work of Binet, who developed the first scale that correlated importantly with the criteria considered to indicate intellectual or scholastic ability [Weiss and Betz 1973]. It is thus that a new period in the history of psychometrics began about 1904 with important contributions by Binet in France and Spearman in England.

Alfred Binet (1857-1911), whose father and grandfather were physicians and whose mother was an artist, took a degree in law and began to study medicine. At Sâlpetrière, the

mental hospital where Charcot did his mental teaching, Binet soon became primarily interested in psychopathology and psychology. Working mainly with children, he began a long series of experiments in memory, movements, sensation, perception, illusions, suggestibility, comprehension, and aesthetics [Binet and Simon 1905].

In October, 1904, the Minister of Public Instruction in Paris named a commission charged with the study of measures to be taken for insuring the benefits of instruction to defective children. Binet was a member of that commission, which decided that no child suspected of retardation should be eliminated from regular school without proper pedagogical and medical examination. Binet and his associate, Theodore Simon, began working earnestly to develop a measure of intelligence to overcome the ills of subjectivity that plagued diagnosticians of mental retardation at the time [Binet and Simon 1905].

Diagnosing retardation, as viewed by Binet and Simon [1905, p. 40], employed three methods:

1. The medical method, which aims to appreciate the anatomical, physiological, and pathological signs of inferior intelligence.

2. The pedagogical method, which aims to judge intelligence according to the sum of acquired knowledge.

3. The psychological method, which makes direct observations and measurements of the degree of intelligence.

Pursuit of the psychological method resulted in the establishment of a measuring scale of intelligence, the Binet-Simon Scale.

Binet and Simon developed the first psychological tests, consisting of separate items, chosen systematically in relation to difficulty level and outside criteria, and published with careful instructions for administration and interpretation. The 1905 scale began with simple coordinating tasks and progressed through more complicated exercises involving reasoning and memory. The detailed instructions called for one examiner to test a single examinee in a quiet comfortable setting, free from distractions. Binet and Simon reasoned that a single examiner could best encourage the pupil to respond. They cautioned, however, against the possibility that examiners might unwittingly assist the subject.

In 1908, Binet and Simon published a revision of their intelligence scale. Instead of some thirty tests arranged in order of difficulty, the new scale consisted of fifty-eight tests arranged in age groups, from age three through age 13. This scale was widely adopted in Europe and in the United States. Since then there have been changes in the assignment of particular tests to various age levels and other improvements have followed, but the Binet method remains basically as it was originally described [Weiss 1973].

A major development in connection with the interpretation of results of mental testing was made by a German psychologist, William Stern (1871-1938), who pointed out that retardation of a certain amount had different meaning at different ages. Accordingly, he suggested that the mental age be divided by

the chronological age thus yielding the "mental quotient." Through refinements to this idea such as the removal of decimal points and methods to insure uniformity about the mean and standard deviation, the mental quotient has become the intelligence quotient or I.Q. in common usage today [Stern 1912].

Although Binet had knowledge of the statistical methods that were being developed during the time of his work on individual differences, he used them infrequently. One of Wundt's students, Charles Spearman (1863-1945), utilized correlational concepts in formulating a number of principles that have become important parts of psychological test theory. Under Wundt, Spearman's principal endeavor was experimental psychology, but he also found time to study the works of English statisticians Pearson and Yule.

While serving a period of military service in Guernsey and England during the Boer War, Spearman collected data on pupils in a village school in Hampshire. He produced two important papers from the data: "The Proof and Measurement of Association between Two Things" [Linden and Linden 1968] and "'General Intelligence,' Objectively Determined and Measured" [Spearman 1904]. The first paper introduced the "correction for attenuation" and eventually led to the development of the concept of test reliability; the second presented the core of Spearman's "two-factor" theory of intelligence, and eventually led to the development of methods

for locating a general factor underlying a group of tests [Linden and Linden 1968].

Spearman's early work depicted a movement toward more precise methods of standardizing tests and of calculating their results. In his 1904 article on intelligence, Spearman criticed previous tests, outlined major problems that should be studied, and indicated the techniques by which these problems might be attacked. Previous work on mental testing was criticized on four points: (1) investigators had failed to use precise quantitative expressions to represent the degree of correlation between tests, or between tests and other measures; (2) the previous work did not include calculations of the probable error of the correlation; (3) certain irrelevant or falsifying factors that might produce misleading correlations were not eliminated; and (4) errors in observation were not taken into account. In short, Spearman emphasized the importance of employing precise measures of calculation [Linden and Linden 1968].

In calling attention to the complexity of factors that affect a correlation coefficient, Spearman proposed that certain extraneous factors may influence the result, such as kinship between individuals who are tested, differences or likenesses of the social class or age of such individuals, and differences in attitudes or abilities. Spearman thus developed a formula to show what the correlation between factors would be if measurement errors were eliminated; the

formula is known as the "correction for attenuation" [Gullikson 1950].

Spearman devoted much attention to the development of a theory of "general intelligence." He rejected the faculty psychology that had evolved from the mental philosophers and formed much of the work of the experimental psychologists. He developed a statement about the nature of intelligence as a "common central factor" that participates in all sorts of special mental activities [Terman 1916]; thus tests could be deduced to measure it. Spearman also developed mathematical procedures that could be used to test the theory.

Spearman designated the common factor, general intelligence, with the mathematical symbol "g." He hypothesized the existence of specific factors that he labeled "s's" [Tuddenham 1962]. The correlational procedures he used to support his theory of general intelligence marked the beginning of factor analysis, a method for summarizing the correlations among a large number of measures in terms of a smaller number of factors [Dunnette 1966].

Sir Cyril Burt (1883-1971) developed the verbal analogy which has become a popular means of measuring intelligence. Burt, an early associate of Spearman's, made contributions in applied psychometrics through development of new instruments, by conducting testing programs, and by refining methods for finding the factors underlying mental tests [Dubois 1970].

Lewis Terman (1877-1956) was born and reared in south central Indiana. He studied psychology at Clark University under G. Stanley Hall and E. C. Stanford. A great admirer of Galton, Terman became interested in using mental tests in the study of precocious children. He received his doctorate in 1905 with a thesis on mental tests [Linden and Linden 1968]. In 1910, after teaching a few years as Professor of Child Study and Pedagogy at the Los Angeles State Normal School, he joined the faculty at Stanford University where he remained until his death. Shortly after arriving at Stanford, Terman began his work on revision of the Binet-Simon Scale.

H. H. Goddard had translated the 1908 Binet-Simon Scale into English and introduced it at the Vineland, New Jersey, Training School for mentally retarded children. The Scale was considered inappropriate for children. The result of Terman's work to revise the Scale was published in The Measurement of Intelligence [Terman 1916]. Although the final instrument bore little resemblance to its predecessors, Terman chose to name it the "Stanford Revision of the Binet-Simon Scale of Intelligence." Measurement of Intelligence provided a major impetus to the use of mental tests in America [Linden and Linden 1968].

E. THE DEVELOPMENT OF GROUP TESTS

Arthur S. Otis (1886-1964), one of Terman's graduate students, introduced a unique innovation in the concept of

mental measurement. Otis approached Terman in 1912 with the idea of tests that would accomplish the same measurement as the Binet-Simon but that could be administered to a group of people simultaneously. Terman agreed and for five years Otis worked to develop test items that could be administered to groups. The items were selected and arranged into a formal scale and first standardized on a representative sample of the Stanford population in 1917. The scale was named the Otis "Absolute Point Scale" [Linden and Linden 1968].

The Otis scale was essentially a battery of tests containing two complete sets of test items but with different specific content. The ideas for appropriate test items were devised from a variety of sources although primarily from Terman's Stanford-Binet. Otis' duplicate sets of tests included items relevant to spelling, arithmetic, synonym-antonym, proverbs, disarranged sentences, relations, geometric figures, following directions, and narrative completion. Geometric designs were attributed to A. R. Ableson. Completion items involved the concepts of Ebbinghaus, G. M. Whipple, Terman and others. Synonyms and antonyms were unique to the Otis scale. Otis had not published his test when the United States entered World War I in 1917, but it became invaluable as the prototype for large-scale testing in the military [Linden and Linden 1968].

F. WORLD WAR I AND MENTAL TESTING

The onset of World War I created a pressing need for a mental test to identify men who were unfit for service because of a lack of intelligence, to sort out those who were more intelligent for further training, and to provide more nearly balanced units.

When war came, Robert M. Yerkes (1876-1956) was the president of the American Psychological Association. In order to deal with the scientific problems of a psychological nature, the National Research Council organized the General Committee on Psychology for the purpose of organizing and supervising psychological research and service in the war effort. Yerkes, who was appointed chairman of this committee, presented a detailed report of the committee's formation to the American Psychological Association (APA) in December, 1917 [Yerkes 1917]. James McKeen Cattell, G. Stanley Hall, and E. L. Thorndike represented the National Academy of Sciences; Raymond Dodge, S. I. Franz, and G. M. Whipple represented the APA and C. E. Seashore, J. B. Watson, and R. M. Yerkes represented the American Association for the Advancement of Science.

Twelve sub-committees originally were recommended by the Special Meeting of the Council of the American Psychological Association; however, the General Committee organized eleven to deal with either psychological problems per se or problems involving psychological aspects. These committees worked

intensely in 1917 and presented plans to the War Department which were revised and improved and subsequently approved in the summer of that same year [Yerkes 1917]. Psychological services to the war effort grew out of these plans. The Committee on Classification of Personnel in the Army developed and introduced throughout the Army methods of classifying and assigning enlisted men in accordance with occupational and educational qualifications and methods of rating officers for appointment and promotion.

The significance of the contribution of this early work, applying psychological methods to selection and placement in the military, is stated by the president of the American Psychological Association.

> The services of this committee, to the work of which the War Department dedicated nearly a million dollars, ultimately touched, and more or less profoundly modified, almost every important aspect of military personnel work [Yerkes 1920].

The types of services initially provided by the committee are summarized as follows:

- Research of important pertinent literature in the field.
- Psychological examination of recruits.
- Selection of men for tasks requiring special aptitude.
- Investigation of psychological problems in aviation.
- Work on psychological problems of incapacity, including shell-shock, reeducation, etc.

- Work on psychological problems of vocational characteristics and vocational advice (combined with incapacity above).
- Investigations into psychological problems of military recreation.
- Work on education and training in the military.
- Work on problems of motivation, emotional characteristics, acoustics, and vision.

The work of the committee, organized for the psychological examination of recruits, is most important to psychometrics. Yerkes chaired this committee.

Yerkes assembled the committee on the examination of recruits in May, 1917, at the Training School, Vineland, New Jersey. The Committee decided that psychological tests offered the best possibility for practical service to the military and agreed that group testing was the best method. Upon developing initial tests, the committee tried them at four different military locations. Five groups of three men each surveyed approximately four thousand soldiers and compared the results. The committee felt that the results justified the belief that the group testing method would be serviceable to the Army [Yerkes 1917].

Following a brief period of small-scale trials of test items and forms, the celebrated Army Alpha test was instituted and named "Group Examination Alpha." The test consisted of eight subtests as follows: oral directions, arithmetical reasoning, practical judgment, synonym-antonym, disarranged sentences, number series completion, analogies, and information.

In addition to the Army Alpha, a Group Examination Beta was created to test foreign (lacking English language skills) and illiterate (presumably below fourth-grade level) subjects. The original form of the Beta examination consisted of 15 tests, most of which were essentially Alpha tests that were translated into pictorial form so they could be pantomimed or demonstrated as opposed to requiring written or oral directions and responses.

The Army testing program, which was under the direction of Yerkes throughout World War I, was the first large-scale use of intelligence tests; 1,726,966 men were examined. Such magnitude coupled with other successes served to change the way the field of psychology would be viewed. Psychology, which was previously considered to be largely an academic discipline, began to be viewed as a profession [Dubois 1970].

In 1918, Otis published the group test that had been the primary model for the Army Alpha. The original Otis test, the Absolute Point Scale, was renamed the "Otis Group Intelligence Scale, Advanced Examination." Designed for use in grades five through 13, this Otis test rapidly gained extensive application.

Since the widespread recognition and acceptance of intelligence testing brought about by the Army testing program, attempts to measure wider varieties of mental factors have followed, including the measurement of academic achievement, special aptitudes, interests, and personality characteristics. For our purposes here, we will turn to developments

in testing specific to military selection and placement, reporting from the time of the Alpha and Beta tests through the current state of the art in military testing.

III. MILITARY MENTAL TESTING BEYOND 1917

The purposes to which tests have been applied are wide-ranging. Most of us are familiar with various achievement tests and aptitude tests. We have undergone numerous examinations aimed at measuring retention levels of course materials presented in our schools. We are also familiar with tests which measure our skills such as driver's license tests and tryouts for sports teams. There are a number of college and professional entrance examinations such as the Scholastic Aptitude Test and Graduate Record Examination. The military services have also carried out an aggressive and ambitious testing program over a broad range of abilities. The discussion which follows, however, is mainly concerned with selection, classification, and placement of personnel at the entry level. It is recognized that these entry-level tests serve a variety of functions other than simply that of qualifying for entry. Some of those functions will be mentioned briefly as we look at testing after World War I.

A. INITIAL GROWTH OF TESTING IN SOCIETY

It is safe to say that prior to the 20th century, psychology was struggling for survival. It had gained neither a respectable scientific reputation nor a marketable professional product. From this relatively obscure position, psychology rose to scientific and professional prominence within twenty-five years largely on the strength of

the military testing program of World War I [Marks 1976].

Indeed, psychology's moment arrived quite dramatically. In 1919, a psychologist, James R. Angell, became chairman of a new government organization of natural scientists, the National Research Council (NRC). We have already noted Robert M. Yerkes' rise to prominence in military intelligence testing. He was appointed chairman of an important NRC division, the Research Information Service, and a separate NRC division was set up for psychologists which functioned as a committee in the Division of Medical and Related Sciences [Samelson 1977].

Yerkes' enthusiasm for the success of the Army program prompted him to campaign for wider use of intelligence tests. In a number of speeches, articles, and books, Yerkes and others publicized their work with tests in the military, its methods, and the results. One of the greatest mediums through which the Army testing program gained its widespread acceptance was the final report published in 1921 with Yerkes as editor. This report, known as the "Army Report," sparked a substantial surge of intelligence tests [Pastore 1978].

Among the promises fostered by the psychologists were the ability to regulate human behavior -- human engineering -- and support of meritocracy -- impartial treatment of everyone [Herrnstein 1971]. The Army testing program gave psychologists the unprecedented authority to select and control a large population; that is, the science of human engineering

was legitimized. As regards meritocracy, psychological tests were objective and impartial since the same tests were given to everyone and objective standards were applied to scoring. Additionally, the test items were thought to measure innate abilities exclusive of environmental influences [Marks 1976, p. 5].

James McKeen Cattell, whom we discussed in the previous chapter, recognized the significance of initial psychological testing successes. In order to ensure that psychologists capitalized on this new-found support from society, he founded the Psychological Corporation. W. V. Bingham, G. Stanley Hall, W. B. Pillsbury, Charles H. Judd, R. Dodge, Lewis Terman, E. L. Thorndike, and Yerkes joined him as initial directors [Marks 1976, p. 6]. The idea behind the corporation was to provide for further psychological research. By combining their money, members hoped to generate business by selling their services; their chief salable service was mental testing.

Inspired by the initial success of the Army program, the General Education Fund gave a grant of $25,000 of Rockefeller money to the NRC for the construction of a group intelligence test for school children. This test, the National Intelligence Test (NIT), was given to millions of children in the twenties. In less than a year, 575,000 copies sold. From 1922 to 1923, 800,000 additional copies sold. In 1923, one firm that dealt in intelligence tests sold 2,500,000 tests, and by 1926, there were 30 companies dealing in intelligence tests [Freeman, 1926].

The Carnegie Foundation was also heavily involved in the testing movement. Between 1905 and 1951 the Rockefeller and Carnegie Foundations contributed $6,424,000 toward testing [Marks 1976]. This philanthropic endowment was not engaged in for purely "goodness of heart" reasons, however. According to Marks [1976, p. 7],

> ...philanthropic foundations had particular reasons for participating in defining individual differences and in promoting intelligence testing. While it was often argued that the two largest philanthropic foundations, the Rockefeller and the Carnegie, were not promoting industrial interests but the general welfare, the evidence indicates the contrary. ...The system needed to preserve and educate a talented elite, to secure a means of selecting workers, to assure social control through fitting individuals into their places in society, and to provide a rationale for the unequal distribution of wealth. The psychological profession and philanthropic foundations aided industrial capitalism nicely in securing these ends.

One goal of testing was to identify and educate the intellectual elite. Other functions included identification of intellectual deviance and social controls. These functions contributed to a fairly widespread movement for sterilization of the mentally deficient and restriction of mentally inferior immigrants [Kamin 1974]. Thus, psychological testing was both heralded by its advocates and criticized by its doubters. Nevertheless, it grew rapidly.

1. Army Results

Following the successful application of psycholigical methods -- that is, the demonstrated efficiency of applied psychological science -- the focus shifted to analysis of the findings contained in the Army testing data. Apparently,

there had been some suspicion that the purpose of the work the psychologists were doing was for the collection of scientific data rather than for helping the Army. Some of the information collected, such as home town, state or county of birth, were interpreted by the Army as evidence that its interests and those of the psychologists were different in some significant respects [Samelson 1977]. In addition to this general lack of trust, the data revealed that the average mental age of white recruits was 13.08 years [Pastore 1978]. Further, since the sample was sufficiently large, the 13-year mental age was regarded as reasonably estimating the average mental age "of adult white Americans in the population at large" [Pastore, p. 317]. This finding was particularly shocking since twelve years was considered to be the upper limits of feeble-mindedness [Samelson 1977].

The nation was shocked over the report of mental intelligence. Indeed, approximately three percent of Americans, according to the Army Report, had mental ages below ten years. H. L. Mencken was so pessimistic concerning the results that he wrote that "...a new breed of man was being spawned in the Western Hemisphere, 'Boobus Americanus'" [Reisman 1966]. The public outcry proved to be too much for the Army to tolerate, and a political shift was made away from testing in the military.

Following the Armistice in 1918, the new peacetime Army quickly dropped the intelligence testing of its soldiers. The War Plans Division was apparently concerned with the

well-being and usefulness of recruits and did not want men of inferior intelligence to be officially identified by test scores. It was feared that the men could become objects of public ridicule and the butt of practical jokes. The effects on morale and efficiency would no doubt be too costly. There was a general resistance to the use of psychological assessment as a selection method, and military mental testing subsequently dropped from the literature until World War II [Wilkins 1972].

2. World War II

The Army, Navy, and Army Air Force (which later became the U.S. Air Force) each had extensive psychological testing programs during World War II. In 1939, when it seemed that war was about to break out in Europe, a Personnel Testing Section was established in the Office of the Adjutant General of the Army. The main emphasis was on the development of instruments to facilitate the classification of recruits and to aid in making job assignments [Goodenough 1961]. The Army General Classification Test (AGCT) was devised as principally a modernized revision of the Army Alpha. The AGCT at first utilized a general overall score, but eventually it had four-part scores as well: reading and vocabulary, arithmetic computation, arithmetic reasoning, and spatial relations. The psychologists who produced the AGCT specifically chose not to label these tests as IQ tests, to make no reference to mental age, grade levels, or innate mental abilities. They sought to avoid mental-age controversies

such as those that followed World War I [Bingham 1941]. Table 1 gives a fairly comprehensive list of Army tests of World War II.

TABLE 1

Tests Developed and Used by U.S. Army in World War II

Classification Tests

- General Classification Test
- Non-Language Test
- Visual Classification Test
- Higher Examination
- Officer Candidate Test
- Women's Classification Test (Mental Alertness Test)
- Army Information Sheet (minimum literacy test)

Aptitude Tests

- Mechanical Aptitude Test
- Clerical Aptitude Test
- Radiotelegraph Operator Aptitude Test
- Code Learning Test
- Battery of Tests for Combat Intelligence
  - Identification of Aerial Photographs
  - Map Identification
  - Route Tracing
  - Battle Maps
  - Perception of Detail
  - Map Reading
  - Map Orientation

Educational Achievement Examinations

- Algebra
- Arithmetic
- English Grammar and Composition
- French
- General History
- German
- Inorganic Chemistry
- Physics
- Plane and Solid Geometry
- Spanish
- Trigonometry
- United States History
- Combined Algebra, Trigonometry and Geometry

Trade Knowledge Tests

General Automotive Information Test
General Electricity and Radio Information Test
General Radio Information Test
Driver and Automotive Information Test

Warrant Officer Examinations

About 30 technical examinations

Army Specialized Training Program Tests

Army Specialized Training Program Test
(achievement tests in each subject taught under the program are under construction)

Source: Staff, Personnel Research Section, Classification and Replacement Branch, the Adjutant General's Office in Science, v. 97, p. 473, 28 May 1943.

Not much literature was available on the Navy's testing program although the Navy did employ a general classification test. A number of tests were developed for use in the selection of naval officers, pilots, instructors, and candidates for training in specialized skills [Davis 1943].

Because the Navy had remained relatively small during World War I, it had not been necessary at that time to establish an extensive personnel selection and placement program. The requirement to procure, classify, train, and assign four and one half million officers and men to man the ships and shore stations in World War II called for modern personnel techniques. The Navy sought assistance from the civilian sector and reorganized its internal structure in

order to apply the most up-to-date techniques of that time to its personnel problems. The result was the establishment within the Bureau of Naval Personnel of a unit of personnel trained in psychology which became the Test and Research Section [Stuit 1947].

The purpose of the Test and Research Section was to develop tests and carry on research studies designed to assist in selecting, classifying, and training officers and enlisted personnel from the time they were examined for admission to the Navy, through their indoctrination or basic training and specialized technical preparation, until they were satisfactorily assigned to duty at sea or on shore.

In December, 1941, the following tests were in general use in the Navy: General Classification Test, Mechanical Aptitude Test, Arithmetic Test, English Test, Spelling Test and Radio Aptitude Test. These tests were criticized because they were not accurately placing people in the proper training schools. For example, the tests did not discriminate between good candidates for radioman training and good candidates for storekeeper school [Stuit 1947]. Subsequently, the Navy followed two paths toward trying to improve its testing program, one within the Training Section and the other on a contract basis under the National Defense Research Committee.

One of the requirements as the war progressed was to be able to assign a large number of people to technical training programs so that there would be the least possible

amount of attrition. In order to meet the challenge, the Navy expanded its Bureau of Naval Personnel. Several psychologists were commissioned to assist with testing. Work was thus begun to revise and update officer and enlisted tests [Stuit 1947].

The contract program was a joint effort among the Army, Navy, and the National Research Council. Bray [1948] has given a general account of this work done under the Applied Psychology Panel of the National Defense Research Committee. The subsequent Basic Test Battery consisted of a (1) General Classification Test, (2) Reading Test, (3) Arithmetical Reasoning Test, (4) Mechanical Aptitude Test, (5) Mechanical Knowledge Test (Mechanical Score), and (6) Mechanical Knowledge Test (Electrical Score). The experimental forms of three tests in the new Basic Test Battery were developed and administered by the end of March, 1943, to obtain data for item analyses. On the basis of data analyses from six naval training centers, the tests were revised and printed in book form ready for routine administration in June, 1943.

Three forms of the tests were in use during the remainder of World War II. New tests of clerical aptitude, spelling, and radio-code aptitude were added to the battery. The general tests of the battery were found in four test booklets as follows:

Book 1. General Classification Test

Book 2. Reading Test and Arithmetical Reasoning Test

Book 3. Mechanical Aptitude Test

Book 4. Mechanical Knowledge Test (Mechanical and Electrical Scores)

The special aptitude tests (clerical, spelling, and radio code) were issued in separate books [Stuit 1947].

The Army Air Force developed an elaborate system of tests and measurements. Most notable among these tests was the Army Air Force Qualification Examination which was given between 1942 and 1946. The examination was given to more than a million high school and college graduates and appeared in some 17 separate forms containing 2,910 different items. Numerous tests were developed and validated including vocabulary, reading comprehension, contemporary affairs and aviation, judgment and logical arithmetic ability, other forms of reasoning, and perceptual abilities [Davis, 1949].

One of the most significant advances from World War I to World War II in the field of military psychology was in the area of classification. The services employed personnel specialists who would personally interview each recruit. These specialists, during informal discussion, would collect a vast array of data such as previous work experience, educational background, specialized interests, and qualities of leadership. These data, together with test scores, enabled commanders to make more informed decisions on the training,

placement, and utilization of military personnel [Davis, 1943]. Figure 1 shows a typical flow chart of the Army's classification system. Each service was similar in many respects. Over the years ensuing after World War II, the Basic Test Batteries employed by the respective services continued to form the foundation of their selection and placement programs.

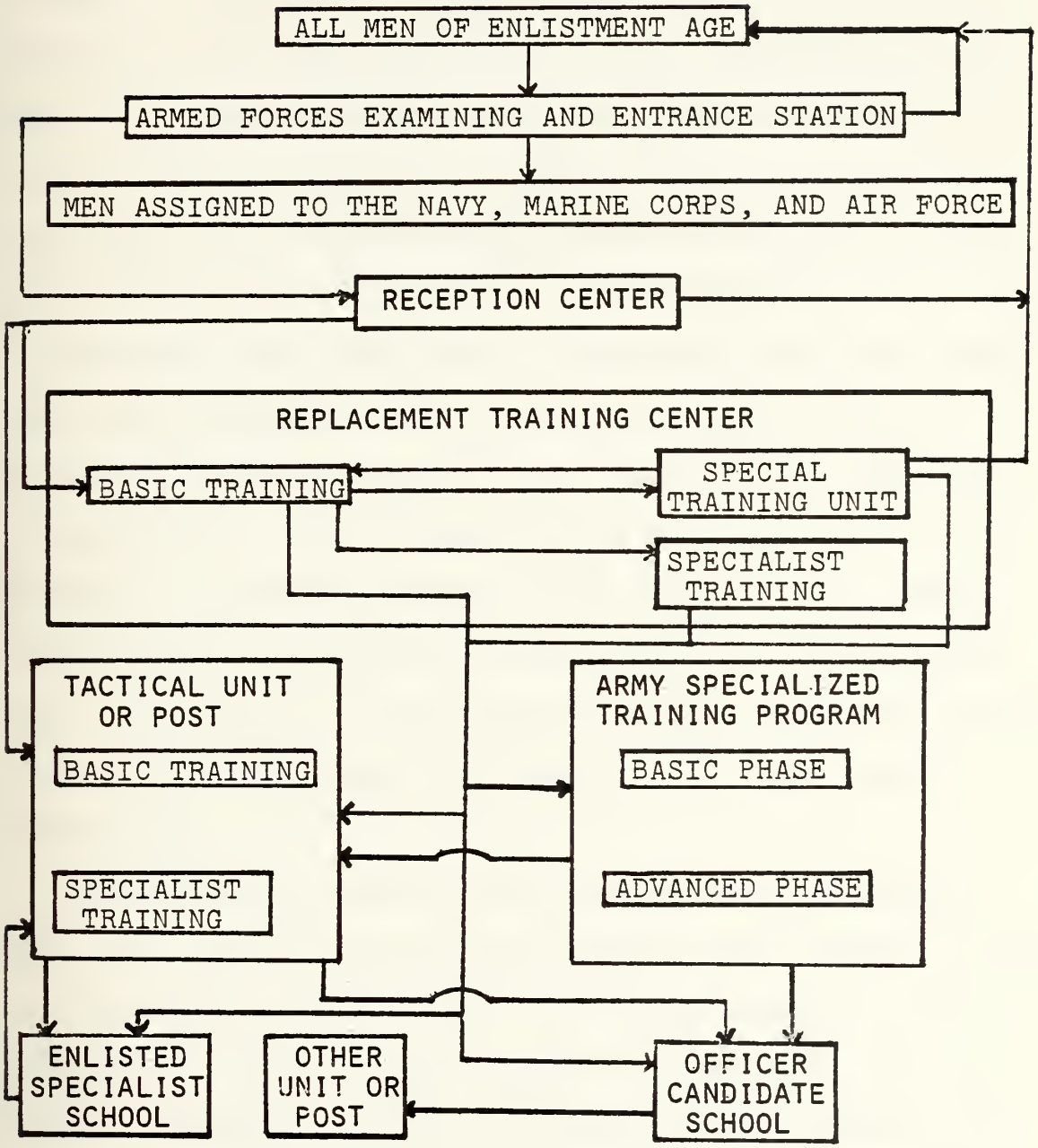
B. AFQT AND ASVAB-1

In 1948, in response to congressional legislation (Public Law 759, 80th Congress, 1948), the Army, Navy, and Air Force provided technicians to prepare an armed forces test for screening recruits for the three services and allocating quotas under the Selective Service. Major emphasis was placed on a definitive acceptance/rejection cutting point [Brandt 1949]. The result of the collaborative effort was the Armed Forces Qualification Test (AFQT).

The AFQT was administered to all potential enlistees, both voluntary and Selective Service applicants, and the principal aim was to predict overall trainability. Testing for specific aptitudes as a basis for classification of enlisted personnel for training and jobs continued to be carried out by the individual services using their respective test batteries. The Marine Corps used the Army tests in screening and classification [Bayroff and Fuchs 1970]. The AFQT was composed of 100 questions, or items, equally divided among word knowledge, arithmetic reasoning, spatial

FIGURE 1

FLOW CHART-ARMY CLASSIFICATION SYSTEM



perception, and knowledge of tool functions [Maier and Fuchs 1973].

The AFQT was introduced in 1950, and research produced improved forms of the test from time to time over the next 20 years. It has been estimated that over 1,000,000 potential failures -- those who would have proven untrainable in the military -- were screened out in that same 20-year period. In 1970, particularly, over 55,000 potential failures were identified, thus avoiding annual accessioning and training costs estimated at $330 million [Maier and Fuchs 1973, p. 51].

The basic test batteries of the several services contained tests which appeared to be similar in content although differing in format, length, difficulty pattern, and other characteristics. For example, tests of verbal ability and arithmetic reasoning appeared on all the services' tests. The question was repeatedly raised: Why not have a single test to be used by all the services rather than three different tests all of which appeared to measure the same aspects of trainability?

In addition to the question of similarity of content, there came a practical problem related to the testing of high school seniors as part of the recruiting programs of the Army, Navy, and Air Force. For a number of years, the Air Force had been administering the Airman Qualifying Examination in a large number of high schools. Test scores were made available to school counselors for use in student guidance,

as well as to Air Force recruiters. When the Army and Navy sought to test in the high schools, each with its own test battery, the additional testing time required brought considerable resistance from the schools. If testing in the high schools for recruiting purposes was to be continued, the testing time required would have to be reduced. A logical solution was for all the services to use the same battery [Bayroff and Fuchs 1970, p. 2].

The Manpower Management Planning Board, of which the Assistant Secretary of Defense (Manpower and Reserve Affairs) was chairman, requested the research representatives of the services to review the technical problems involved in developing a single test battery for use by all the services. The battery was to serve the following purposes:

1. Testing high school seniors.
2. Establishing mental qualifications for enlistment and induction.
3. Selection of enlistment applicants for particular occupational or training systems.
4. Classification and assignment.

The recommendations made by the research representatives were that such a design was feasible [Bayroff and Fuchs 1970, p.2].

In February, 1966, the Assistant Secretary of Defense directed the services to begin development of a common aptitude battery to be given to high school seniors.[1]

---

[1]Memorandum for the Undersecretaries of the Military Departments from the Assistant Secretary of Defense (Manpower), Subject: "Development of a common aptitude battery," dated 3 February 1966.

He also directed that the testing time be no longer than two and one half hours. All four services participated in the study, the Army as lead service having major responsibility. Through a series of sampling tests conducted on in-service personnel by each of the services, the following was accomplished: interpretation of data (each service provided punched cards containing scores on test batteries to the U.S. Army Behavior and Systems Research Laboratory for statistical processing); identification of the interchangeable tests; selection of items for an abbreviated test; and standardization of the abbreviated tests. The resultant test was called the Armed Services Vocational Aptitude Battery, Form 1 (ASVAB-1) [Seeley, Fischl, and Hicks 1978].

ASVAB-1 was used for pre-service testing in high schools from 1968 until it was replaced by the improved parallel forms ASVAB-2 and -3 in January 1973. The U.S. Army Research Institude for Behavioral and Social Sciences Technical Paper 289 of February 1978 contains a summary of ASVAB-1 and the rationale for development of ASVAB-2 and -3. ASVAB-1 consists of the following nine tests:

1. Coding Speed
2. Word Knowledge
3. Arithmetic Reasoning
4. Tool Knowledge
5. Space Perception
6. Mechanical Comprehension

7. Shop Information

8. Automotive Information

9. Electronics Information

C. DEVELOPMENT OF ASVAB-2 AND -3

Immediately upon introduction of the ASVAB into high schools, work commenced on the development of two successor forms. The objective was to develop a pair of parallel forms which would be comparable but superior psychometrically to the form then in use in the schools. ASVAB-1 contained some tests in which the mean item difficulty was other than the most desirable. In addition, parallel forms were required for retest purposes. The research steps to develop the two new forms required (1) preparation of test items, (2) field administration to obtain empirical data concerning the items, (3) selection of the items to comprise final test forms, and (4) a second field administration to derive norms, intercorrelations, and test reliability coefficients [Seeley, Fischl, and Hicks 1978].

Two hundred new test items of each of eight types of content were administered in Armed Forces Examining and Entrance Stations (AFEES) to several national samples of Selective Service registrants stratified to represent the population of young men of military age. The total sample consisted of some 4,000 cases, 18% of whom were blacks, and requisite statistics of item difficulty and homogeneity were obtained. Using these statistics, items were assembled into

two parallel 25-item forms of each of these eight content types. New forms of the Coding Speed test were generated, and when these were added to the item-analysis-based tests, there were two entirely new nine-test batteries [Seeley, Fischl, and Hicks 1978].

These batteries were subsequently administered to several additional stratified national samples of Selective Service registrants, one form to a sample. A total of 3500 cases, at 13 AFEES, was utilized for this administration. From this administration, percentile and Army Standard Score norms were developed; and test reliability coefficients, intercorrelations, and other characteristics of the new batteries were derived. The new forms became ASVAB-2 and -3 which were used in the high school testing program. ASVAB-4 was developed as a back-up to ASVAB-2 but was never released.

D. ASVAB-5, -6 AND -7

In early 1974, DOD directed that the services move jointly and expeditiously toward use of a common aptitude battery for classification and placement of enlisted personnel. Recall that the services employed their respective classification tests for this purpose up to that time. The Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs) directed that the ASVAB be redesigned to satisfy enlistment production requirements of all the services, with high school usage being a secondary consideration. The previously developed ASVAB forms lacked some characteristics

which could satisfy enlisted production requirements for all the services.

Jensen, Massey and Valentine [1977] provide a complete description of the development of ASVAB-5, -6 and -7. A pool of 2700 items was assembled for experimental tryout and item analysis preparatory to final item selection. Following essentially the same pattern as the previous ASVAB development, the item pool was administered to samples of 500 to 600 cases at basic training centers and at AFEES. AFEES testing was necessary to obtain representation of cases rejected for enlistment.

Three forms of the battery (5, 6, and 7) were developed from these items. Following further testing and norming of the forms, ASVAB-5, -6 and -7 were delivered for operational implementation as of 1 January 1976. These three ASVAB forms, used until 1 October 1980, were characterized by complete coverage of cognitive materials previously present in the classification batteries of all the services.

In order to satisfy the needs of the services, ASVAB-5, -6, and -7 were expanded from the nine tests in ASVAB-2 to twelve in ASVAB-5; however, a short interest inventory (the Army Classification Test) was also administered as an additional test. ASVAB-6 and -7 were parallel forms of ASVAB-5 and were administered only at AFEES. ASVAB-6 and -7 were placed into use by the AFEES in January, 1976, and use of ASVAB-5 did not begin until July, 1976. Tables 2 and 3

provide a breakdown of the forms of the ASVAB and the content of ASVAB-5, respectively.

TABLE 2

CONTENT - FORM 5 OF ASVAB

Test	Number of Items	Time (Minutes)
General Information (GI)	15	07
Numerical Operations (NO)	50	03
Attention to Detail (AD)	30	05
Word Knowledge (WK)	30	10
Arithmetic Reasoning (AR)	20	20
Space Perception (SP)	20	12
Mathematical Knowledge (MK)	20	20
Electronics Information (EI)	30	15
Mechanical Comprehension (MC)	20	15
General Science (GS)	20	10
Shop Information (SI)	20	08
Automotive Information (AI)	20	10
Totals	295	135

Source: ASVAB Mini-Guide, U.S. Government Printing Office: 1978-752-924 MEPCOM05COUN

TABLE 3

THE TWELVE CONTENT AREAS WHICH APPEAR ON ASVAB FORMS -5, -6, and -7

GENERAL INFORMATION (GI) Measures a portion of a student's developed ability to recognize factual information characterized by the cumulative influences of his or her learning experiences.

NUMERICAL OPERATIONS (NO) Measures an individual's developed ability to rapidly and accurately compute simple number computations.

ATTENTION TO DETAIL (AD) Designed to measure the ability of an individual to perceive simple relationships, to retain these relationships mentally, and to make decisions based upon the relationships involved quickly and accurately.

WORD KNOWLEDGE (WK) Measures verbal comprehension which entails the ability to understand written and spoken language.

ARITHMETIC REASONING (AR) Designed to measure general reasoning. It is concerned with the ability to generate solutions to problems. It is different from Numerical Operations in that the student must construct a solution by some principle in order to solve the given problem.

SPACE PERCEPTION (SP) Measures an individual's spatial aptitude. This infers an ability of an individual to visualize and manipulate objects in space.

MATHEMATICS KNOWLEDGE (MK) Measures functional ability in the use of learned mathematical relationships. Factors measured by this area tend to overlap the areas of numerical operations and arithmetic reasoning. The similarities are in the functions performed. The differences lie in the complexities of the functions.

ELECTRONIC INFORMATION (EI) Measures functional ability in the use of learned electronic relationships. A number of factors appear to be measured by this test: arithmetic reasoning in the form of simple electronic calculations; verbal comprehension in terms of the person's reading level with respect to electronic terminology; and a level of general reasoning is indicated by having the individual make use of electronic principles in order to arrive at the correct answer.

MECHANICAL COMPREHENSION (MC) Measures the ability of an individual to learn, comprehend, and reason with mechanical terms. Even though familiarity with common tools and mechanical relations is a prerequisite, further technical knowledge is not necessary other than that acquired through day-to-day experiences. This test has pictures of mechanisms whose functions call for comprehension.

GENERAL SCIENCE (GS) Measures a level of verbal comprehension in the general area of science. This test was designed to measure a form of reasoning which involves the ability to see the relationship between two factors or scientific ideas. Some arithmetic reasoning may also be involved.

SHOP INFORMATION (SI) Measures the functional ability of an individual who has had experience with and is knowledgeable about the use of a variety of tools found in a shop. In addition, it appears that a level of verbal comprehension is also measured as indicated by the understanding needed of the terminology used.

AUTOMOTIVE INFORMATION (AI) Measures the functional ability of an individual who has had some experience working with automobiles. This test also relies upon an individual's reading ability and verbal comprehension. The questions may pertain to diagnosing malfunctions of a car, the use of (a) particular part(s) of a car, or meaning of terminology.

Source: ASVAB Mini-Guide

The ASVAB is supposed to change annually [Cronbach 1979]. In reviewing ASVAB-5, -6, and -7, Cronbach [1979] stated that there were a number of problems with them. For example, certain subtests of ASVAB-5, he said, were measures of experience and not of talent. Therefore, they would have little value for counseling. "To judge a person as lacking aptitude for trades on the basis of an information test is inappropriate and damaging" [p. 233]. He also states that the items seem not to have been edited well.

ASVAB has not been changed annually, however. ASVAB-5 is still currently used in the high school testing program. ASVAB-8, -9 and -10 were put into use on 1 October 1980 and currently are used for production testing in the four services. ASVAB-8, -9, and -10 were developed in much the same manner as previous forms of the test. Table 4 illustrates the subtests which appear in ASVAB-8, -9, and -10.

TABLE 4

ASVAB 8/9/10 SUBTEST INFORMATION

ASVAB 8/9/10 Subtest

	Subtest Name	Abbreviation
1.	General Science	GS
2.	Arithmetic Reasoning	AR
3.	Word Knowledge	WK
4.	Paragraph Comprehension	PC
5.	Numerical Operations	NO
6.	Coding Speed	CS
7.	Auto & Shop Information	AS
8.	Mathematics Knowledge	MK
9.	Mechanical Comprehension	MC
10.	Electronic Information	EI

E. ADMINISTRATION AND USES OF ASVAB

The ASVAB is administered by the Military Enlistment Processing Command (MEPCOM), with headquarters at Ft. Sheridan, Illinois. The MEPCOM's primary role, as processing agent for all applicants seeking to enter military service, is carried out at the local level by the AFEES. In the Continental United States there are 65 AFEES, one in San Juan, Puerto Rico, and another in Honolulu, Hawaii, and substations in Alaska and Guam. All AFEES commanding officers report to MEPCOM via one of three sector headquarters:

Eastern (Ft. Meade, Maryland), Western (Oakland, California), and Central (Ft. Sheridan, Illinois) [ALL HANDS 1980].

MEPCOM was established in 1976 to combat some of the problems created by the termination -- placing in standby status -- of the draft in 1973. Prior to 1973, the majority of people processed at AFEES were Army applicants, and the AFEES were under the jurisdiction of the Commanding General of the U.S. Army Recruiting Command (USA-REC) [ALL HANDS 1980].

One of the problems following the introduction of the all-volunteer force (AVF) was the realization that people might cheat to gain entry into the service. Additionally, with the new emphasis placed upon attainment of recruiting quotas, recruiters were tempted to cheat as well. MEPCOM was created to provide a greater measure of quality assurance in all areas of enlistment processing. Control of administration of the ASVAB was one of the major changes.

The ASVAB examination is given separately from the recruiting function at an AFEES or at one of 750 mobile-examination-team (MET) test sites around the country. If a student takes the test in a high school and passes, his or her scores are valid for two years. After graduation, the student need not retake the test in order to enlist.

Once an applicant for service passes the ASVAB and a physical examination, he or she will meet with the recruiter to discuss available service options. Recruiters are aided by the ASVAB by the ability to look at composite scores of the content areas and thereby assist the prospective

service member in selection of a service specialty. Table 5 illustrates the type of composites referred to here. Fischl and others [1978] provide a report of a study of the validity of the ASVAB for predicting performance in service technical training schools.

As mentioned above, the ASVAB is currently providing the testing function through which the armed services select, classify and place their recruits. The following chapter will look at some of the criticisms of testing and briefly discuss some of the current issues facing psychological testing.

TABLE 5

NAMES, DEFINITIONS, AND INTERPRETATIONS OF THE
FACTOR ANALYTICALLY DERIVED ASVAB-5 COMPOSITES

Symbol	Name	Definition	Interpretation
VE	Verbal Ability	WK + GI + GS	Measures knowledge of words, and the ability to understand materials and deal with verbal concepts. The composite is a combination of the scores on the Word Knowledge, General Information, and General Science tests.
AQ	Analytic, Quantitative Ability	AR + MK	Measures reasoning abilities as well as those relevant to understanding quantitative concepts. The composite is a combination of the scores on the Arithmetic Reasoning and Mathematics Knowledge tests.
CL	Clerical Ability	3AD + NO	Measures speed and accuracy in using letters and numbers. These are abilities relevant to clerical types of activities. The composite is a combination of the scores of the Attention to Detail and Numerical Operations tests.
ME	Mechanical Ability	SP + MC	Measures understanding of mechanical principles as well as the ability to visualize objects in three-dimensional space. The composite is a combination of the scores on the Space Perception and Mechanical Comprehension tests.

Symbol	Name	Definition	Interpretation
TT	Trade Technical	AI + SI	Measures information relevant to automotive and various shop practices. The composite is a combination of the scores on the Automotive Information and Shop Information tests.
AA	Academic Ability	WK + AR	Measures abilities needed to do well in school and formal types of training. The composite is a combination of the scores on the Word Knowledge and Arithmetic Reasoning tests.

Source: Directorate of Testing (MEPCOM) Technical Research Report 77-4, 1978, p. 8.

IV. ISSUES TO BE FACED BY TESTING

> This bill addresses the growing concern among parents, students, teachers, academic administrators and the general public about the appropriate uses of standardized tests in the admission process for postsecondary schools.
>
> Colleges and graduate and professional schools rely heavily on standardized test results in deciding which students to admit. These examinations have a profound impact on the educational and occupational future of millions of Americans [Weiss 1979].
>
> -- Honorable Ted Weiss
> House of Representatives

Representative Weiss' remarks were made on introducing the Educational Testing Act of 1979, also known as the truth-in-testing bill. There are any number of scholarly works which debate the good and bad points of psychological testing, but the one statement selected here, that of a legislator, demonstrates the growing "publicness" of mental ability measurement. In spite of the continuous review of testing practices, improvements in psychometric theory, critical analyses of test items, and attempts to provide careful instruction on how to interpret and use test results, psychological assessments fall far short of perfection. The testing establishment has, therefore, become vulnerable to public regulation.

If we paraphrase the last sentence of Mr. Weiss' statement as follows: "These examinations have a profound impact on the future military careers of millions of prospective military personnel," we can readily envision the potential impact of testing regulation on military selection and placement assessment.

The ASVAB is vulnerable to criticism and congressional intervention. For example, Congress imposed "quality" goals (referring to performance on the ASVAB) through the 1981 fiscal year Defense Authorization Act [Philpott 1980].

There are a number of issues which have led to the current state of affairs in the testing and measurement industry. Without pointing an accusing finger, it would be safe to say that proponents of psychometrics have failed to counter the arguments against testing to the satisfaction of the increasingly noisy opposition. What we want to do here is examine a few of the issues facing testing in order to form a framework on which to build a strategy for future use of tests in military selection and placement.

Thorndike and Hagen [1977] have identified some current issues in measurement as: (1) testing of minority persons, (2) invasion of privacy, and (3) use of normative comparisons. Other issues -- some similar in various aspects to the three above -- are identified by Green [1978]: (1) multiple choice tests, (2) validity and aggregation, (3) overdependence on test scores, and (4) test fairness or educational disadvantage. And, of course, the military establishment considers recruiter malpractice and the social weal as it concerns providing opportunity for less fortunate individuals (in terms of test scores).

A. TESTING OF MINORITY PERSONS

An issue that has received a great deal of public attention is that of the use of intelligence tests to select and classify members of minority groups whose background and cultures vary from the "white" majority. America is commonly referred to as a "melting pot" society; there are all sorts of subgroups differing in many ways from one another. Generations of black, Hispanic, and other minority people have been measured by intelligence tests that seem to assume everyone grows up exposed to the same middle-class society [Rice 1979]. For example, Green [1978] points out that white students score substantially higher on college admissions tests than do blacks. He relates that national tests of achievement in high school and entrance examinations for law school and medical school show similar patterns.

Such findings are not necessarily evidence of cultural or racial bias. According to Herrnstein, "Tests are not biased simply because some people get higher scores than others, any more than yardsticks are biased because they show some people to be taller than others" [1980, p. 48]. Cleary [1968] defines bias to mean that a test score from a member of group A would be associated with better performance in school than is the same score for a member of group B. People in Group A could reasonably protest that the test is biased against them if it underpredicts their performance. According to Professor Weitzman, on the other hand,

"A bias-free prediction of future performance should make no distinction among individuals who would perform equally well if given the chance" [Weitzman 1980].

The courts are taking an interest in bias in mental testing. The Supreme Court held in Willie S. Griggs, et al vs. Duke Power Company that the use of the Wonderlic Personnel Test and the Bennett Mechanical Aptitude Test was illegal [Huff 1974].

Since 1955, Duke Power Company had required a high school diploma for hiring in any of its departments, except the labor department, and for transfer from its coal handling department to any inside division (operations, maintenance, or laboratory). Before 1965, the company had restricted blacks to the labor department. The company abandoned this policy in 1965, but the completion of high school was made a prerequisite to transfer from labor to any other department. Additionally, employees who wished to qualify for positions in any department other than labor were required to take the Wonderlic Personnel Test and the Bennet Mechanical Aptitude Test. When Griggs and other employees brought a class action suit against Duke Power, the courts ruled for the plaintiff. In the opinion of the Court as delivered by Mr. Chief Justice Burger:

> The facts of this case demonstrate the inadequacy of broad and general testing devices as well as the infirmity of using diplomas or degrees as fixed measures of capability. History is filled with examples of men and women who rendered highly effective performance without the conventional badges of accomplishment in terms of

> certificates, diplomas, or degrees. Diplomas and tests are useful servants, but Congress has mandated the commonsense proposition that they are not to become masters of reality [Huff 1974].

In the interest of science, we want to guard against such biases whether they appear to evolve unwittingly or have been allowed to become institutionalized.

Numerous studies have been made to assess group differences in performance on standardized tests. The 1976, Volume 13 issue of the Journal of Educational Measurement was devoted exclusively to the area of test bias. It seems intuitively obvious that tests designed to measure mental ability which are written in English would discriminate against people for whom English is a second language or, as Rice [1979, p. 33] says, "against black children whose normal 'street English' differs markedly from that customarily used in middle-class society, in schools, and in intelligence tests." However, attempts to identify test items which require specific knowledge that blacks, for example, have no opportunity to acquire have had little success [Green 1978, p. 668]. As Green states, "Items on professionally prepared tests do not tend to favor one group over another, differentially" [p. 668]. Studies thus do not appear always to support intuition.

Professional preparation includes the careful editing, screening, and pretesting of test items. Developers of the ASVAB attempted to overcome problems of bias by sampling the full range of population taking care to "include

representation of women and ethnic minorities in the item analysis samples" [Jensen, Massey, and Valentine 1977].

Further discussion of test bias is beyond the scope of this paper except to reemphasize the importance of the issue. Indeed, the concept of test bias is complicated by the many definitions applied to it [Flaugher 1978]. Jensen [1980] provides an extensive discussion of some aspects of test bias in Bias in Mental Testing.

B. INVASION OF PRIVACY

Many people prefer to have control over information about themselves, to be able to decide what kinds of information they desire to make available to whom and under what circumstances. The question arises concerning what types of information are relevant and predictive of future academic or job performance. Tests which measure job-related skills such as typing or stenography are generally not objectionable to prospective secretaries who view such tests as directly related to the position for which they are applying and society tends to look favorably on driver's license tests as a means to assess the minimal skills required to operate motor vehicles. These are relatively safe areas for measurement. On the other hand, tests which attempt to describe emotional stability, educability, integrity, motivation, etc., delve into areas which many people prefer not to divulge about themselves.

Not only is there concern about the collection of information, but the purposes for which it is obtained and the applications to which it is put are subject to scrutiny as well. If information is collected at the request of the individual for use in a counseling situation, that is one thing. The act of volunteering information implies a willingness to open one's self up for inspection; invasion of privacy becomes a matter of relatively minor concern [Thorndike and Hagen 1977]. However, when the information is being obtained to further organizational goals, then the concern for the rights of the individual mounts.

C. USE OF NORMATIVE COMPARISONS

There is debate over whether individuals should be compared with norms representing the "typical performance of a national, or sometimes a local, sample..." [Thorndike and Hagan, p. 19]. For example, in a critical appraisal of the Scholastic Aptitude Test (SAT), the authors assert, "The undeserved clout of this test is perhaps most evident when disparities occur between students' high school records and their SAT scores. Students who get good high school grades but do poorly on the SAT are called 'overachievers,' as if, during high school, they have somehow transcended their intellectual potential; the implication is that they will not fare so well in college" [Slack and Porter 1980, p. 170]. They go on to state that the term "overachiever" is particularly "specious" when used to disparage the academic potential of

someone who has succeeded in the arena that best correlates with college performance. Nairn and Associates [1980] report on a number of real life cases using disguised names such as Frank Washington, Mark Simons, and Sam Harrison, who had each demonstrated considerable ability but were denied educational opportunities because of standardized test scores.

The implication for the test designer centers around the situational use of information. When should we seek to determine whether one can satisfactorily perform a specific task and when should we ask where he or she stands in relation to others?

D. MULTIPLE-CHOICE TESTS

Test scores which provide the basis for decisions regarding the futures of test takers are determined by machine-scored multiple-choice items. Green [1978] points out that, "Critics argue that multiple-choice items are unfair to the thoughtful, brilliant students who often see more in a test item than was intended, spend too much time on the item, and then select the wrong response." Still others believe that multiple-choice tests are too superficial.

Proponents of multiple-choice items, on the other hand, argue that test items can be developed to measure from low to high levels of ability through the concept of homogeneity, which implies the process of combining a large number of items of similar content but different difficulty. The relationship among such items has been called "factorial

homogeneity" [Dubois 1970]. Additionally, it has been asserted that professionals devote intense effort to preparing test items to measure deep understanding of issues. Herrnstein argues that tests correlate highly with each other even if the items are superficially utterly different [1980, p. 44] so that a strong case may be made for multiple-choice tests. Conversely, critics reviewing the SAT, for example, argue that the multiple-choice score does not add more than a few percentage points to high school grades in the ability to predict success in college [Nairn and Associates 1980, p. 61]. And again the proponent argues, "By aggregation, a reliable homogeneous measure has been constructed out of a lot of unimpressive items. Given enough sows' ears, we can indeed make a silk purse" [Green 1978, p. 666].

E. VALIDITY AND AGGREGATION

Validity simply means how well a test predicts performance. Critics of testing argue that tests do not always predict performance. One study claims the following predictive capability using the roll of dice to predict grade-rank standing within a group.

Percentage of Predictions in Which Random Prediction with a Pair of Dice Is as Accurate as an ETS Test	
SAT	88%
LSAT	87%
GRE	89%
GMAT	92%

Source: Nairn and Associates, The Reign of ETS, p. 65, 1980.

This means, according to the source, that, on the average, for 88% of the applicants (though it is impossible to know which ones) a SAT score will predict their grade-rank no more accurately than a pair of dice.

Now such damaging revelations as these of Nairn and Associates seriously injure the image of testing, albeit that their report is somewhat in error. In a rebuttal to Nairn, ETS states:

> Nairn's claims that the SAT is a poor predictor of performance in college are based on faulty statistics. He uses an incorrect value for the characteristic validity of the SAT (.341) since he mistakenly averages the separate validities of the two parts of the SAT, rather than considering the validity of the whole test (.41). After squaring that coefficient and doing some further arithmetic, Nairn comes to the conclusion above [ETS 1980, p. 2].

What Nairn has done is to confuse the validity coefficients with the number of cases. He arrives at the conclusion by squaring the characteristic validity of the SAT (.341). Then by multiplying the results by 100, he arrives at the number 12. "This he interprets as the number of cases for which a SAT test will provide a prediction more accurate than a random prediction such as throwing a pair of dice" [p. 16]. Taking the complement of this number, he states that rolling dice will be as accurate as using test scores 88% of the time. "Little can be said of Nairn's interpretation except that it is wrong. It is safe to say that no reputable statistics text can be found anywhere that suggests that r^2 indicates the proportion of predictions better than chance predictions" [p. 17].

In countering the argument against validity, Green [1978] states, "Test scores cannot very well predict performance on any particular item of a final examination in the freshman year. Indeed, they don't even predict course grades very well. It is only the very aggregate grade-point average that is well predicted by verbal aptitude measures" [Green 1978, p. 606]. Herrnstein [1980, p. 45] supports the notion that intelligence-test (IQ) scores are not as predictive in college years as they are earlier in the life of a student, and scholastic-aptitude tests correlate highly with these intelligence tests.

F. OVERDEPENDENCE ON TEST SCORES

Perhaps there is too much reliance on test scores in the first place. The credentials of the developers, the painstaking item analyses, the normalization samples, the touted success of tests as predictors of performance, some critics argue, have led people to place entirely too much emphasis on test scores. Selection and placement, entrance to college, and promotion to higher positions are frequently decided by test score alone. Many colleges will not consider applicants for admission who score below a minimum cutting score on scholastic-aptitude tests even though the "total man" picture would predict a high probability of academic success in many cases [Nairn and Associates 1980].

The "total man" picture embodies the many factors which are not measured by intelligence tests such as creativity as

evidenced in certain extra-curricular activities, motivation as depicted in past evaluations, idealism, and experience. Nairn and Associates argue that, to the extent these factors are ignored, people are not "measured" fully. ETS argues that it is not an "arbiter of admissions or 'gatekeeper' to higher education" [ETS 1980, p. 11]; the schools make the selection decisions and in the process can consider other variables as well as intelligence. Colleges are well informed about what ETS tests can do and they know how to interpret the results. Citing a 1972 study, ETS claims that 87.5% of high school graduates who applied to college were admitted to at least one institution by the end of their senior year [p. 12].

Overdependence on test scores can lead to a variety of problems. Military selection cutting scores have prompted critics to charge that "Mental Category Rules Could be a Minority Barrier" [Philpott 1980b]. There is evidence that people of low mental ability who were admitted to the military service through a norming "error" in the qualifying test score are performing adequately [Philpott 1980a]. However, this is in conflict with the study conducted earlier with Project 100,000 [Project 100,000 Report 1969] which clearly indicated that there were differences in performance levels during that experimental program designed to accept lower mental group personnel into the services. Attrition rates were significantly higher (sometimes three to one) for lower mental groups. Perhaps what the critics think is needed is a

more general appreciation of the ability of tests to serve as tools for the decision-maker. The correlation of a test with a performance criterion is a statement about people in general; some may be able to perform well in spite of poor scores while others may score well and exhibit a disappointing performance [Dunnette 1966]; but without testing, what discrimination device would we use to tell who's who? Should test scores be considered in light of other factors? Of course, that's what multiple regression is about.

G. TEST FAIRNESS OR EDUCATIONAL DISADVANTAGE

The National Education Association (NEA), with a membership of 1.8 million teachers, has called for the abolition of all standardized intelligence, aptitude, and achievement tests on the basis that they are at best wasteful and at worst destructive [Psychology Today 1979]. On the other hand, teachers may be complaining because tests show them up as perhaps inadequate. Teachers, parents, and policy makers complain most of all about the use of the tests to track -- and stigmatize -- the low-scoring students in classes that further discourages learning. Conversely, one might ask, "Is not the bench warmer also 'stigmatized'? Is that a reason to discontinue competitive sports?" Psychology Today [September 1979] expresses the editorial view that fairer and more precise tests are attainable and may be on the way.

The issue of fairness of tests has traditionally centered on whether items favor the "white" majority. Some argue that

the larger problem is in the make-up of our educational system. "Poor, urban youths, most of whom are blacks, must develop their potential in inferior schools, among peers who do not value academic achievement, and in a family and community that do not provide much support and encouragement for educational achievement" [Green 1978, p. 669]. As J. B. Watson puts it [in Kamin 1974],

> Where there are structural defects...there is social inferiority...competition on equal grounds is denied. The same is true when "inferior" races are brought up along with "superior" races. We have no sure evidence of inferiority in the Negro race. Yet, educate a white child and a Negro child in the same school -- bring them up in the same family (theoretically without difference) -- and when society begins to exert its crushing might, the Negro cannot compete [p. 178].

Perhaps the answer to fairness lies in recognizing the need for major educational and environmental changes, an idea which armed-services education programs may very well need to consider.

H. SOCIAL DECISION

The FY81 Defense Authorization Bill limits the number of recruits who may be admitted to the Armed Forces who score in mental category IV on the ASVAB. Similarly, as we stated above, colleges, employers, and various institutions place great significance on test scores in determining entry standards and hiring practices. One basic question concerns whether we can afford the luxury of "culling" out what may be very good human resources, those who fail to score well. We must realize that odds for effectiveness are only odds. They

are not absolute assurances -- but respresent the best prediction available. We mentioned the examples cited by Nairn above, but let's take a short look at the military.

Several informed sources have been quoted as saying that the number of Category IV mental-level members of the military services has been increasing over the years since ASVAB was introduced. Table 6 illustrates a study of Category IV by recent fiscal years.

TABLE 6

PERCENTAGE OF CAT IV RECRUITS BY RACIAL GROUPS AFTER RENORMING

	FY77	FY78	FY79
Army	39	38	45
White	30	26	33
Black	59	57	62
Other	53	52	57
Navy	19	17	20
White	15	14	15
Black	40	37	44
Other	37	31	31
Air Force	5	6	9
White	4	5	7
Black	10	11	18
Other	5	8	11
Marine Corps	24	27	27
White	17	19	19
Black	44	47	47
Other	41	40	37
All DOD	26	24	29
White	19	16	20
Black	49	47	52
Other	40	38	40

The chart above shows percentages by racial groupings of recruits who scored in the lowest acceptable mental category, CAT IV, on military entrance tests during the last three fiscal years. In FY79, for example, 15% of the Navy's white recruits, 44% of its black recruits and 31% of its other minority recruits were CAT IV's. Congress this month placed limits on the percentage of CAT IV's that can be recruited in future years. In FY81, CAT IV's for DOD overall must not exceed 25% of all recruits. In FY82, the same limit will be placed on individual services, and in FY83 no service can accept more than 20% of CAT IV's.

Source: Navy Times, October 6, 1980.

Yet testimony regarding the expected decline in performance due to the "inferior quality" of recruits has not been forthcoming. Members of the Committee on Appropriations -- specifically the Subcommittee on the Department of Defense -- have stated that they are told by commanders that the quality of people is as good as it has always been [U.S. Congress 1980]. Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), in testifying before the House Appropriations Committee on April 1, 1980, stated that the services are:

> ...attempting to determine its (the increase in Category IV's) impact...we have launched an effort to learn what difference it does make in terms of performance.
>
> We would have thought that if we really had experienced an increase in Category IV personnel,

then we would have also seen a dramatic increase in attrition from our skill training courses. That doesn't appear to be the case.

...Many senior people with great experience who ought to know what they are talking about say that they have very good quality people, and that they are trainable, and that they are satisfied with them [U.S. Congress, 1980, p. 63].

To which Mr. John P. Murtha, House of Representatives, replied:

The former Chief of Staff said that these are the best troops he has ever had, he would compare these with any troops he ever served with.

...Why do you keep putting out reports which actually degrade the services and the men serving in the services, and then you complain to us that for some reason the American population thinks the people in the American services are not as high quality?

I think I would take the word of the commander over a test score, which you keep changing around all the time.

The testimonies above must be considered in light of the political arena in which they occurred. There is a big debate in progress concerning the all-volunteer force (AVF). It is possible that proponents of the AVF may be more sensitive to and more likely to relate those anecdotal evidences which paint a positive picture of the quality of manpower. There may have been little choice but to take a middle-of-the-road stance on the quality issue because of political considerations. The testimony lacks empirical evidence and quotes no real numbers. Again we have the evidence presented by Project 100,000 which was not mentioned. Rather than speculate on the motivation behind testimony, perhaps we should ask whether it is true that we don't know what the impact of increases in

Category IV's is. Not everyone is undecided, however. The New York Times [1980] cites the ongoing debates over the manpower "quality issue" and discusses what it calls "a severe manpower crisis" exhibited in the slumping intelligence scores of recruits.

There is a social dilemma here which all test developers and users face. Do we arbitrarily categorize people and exclude them from our institutions? Proponents of testing argue that we should continue to use tests and improve our abilities to measure.

According to Green [1978, p. 669], "Objective tests have the same advantages and disadvantages that they have always had. They provide objective, reliable measurement of important skills necessary for academic achievement." Tests currently measure only a few of the qualities needed for performance in education and training schools. "As tests become more accurate and provide indications of more aspects of individual differences, other social problems may be highlighted" [p. 669].

Of necessity, the above discussion has only briefly touched on some of the key issues which must be considered in the development, implementation, and use of tests and their results. The issues are not static; we have dealt with them for more than seven decades. Theorists are working on the issues, and there is hope for great improvement in the next decade. Flexibility and adaptibility will be the key to success in dealing with these issues.

The next chapter will describe a relatively new concept in test administration which may provide the desired responsiveness to changes in testing specific to military selection and placement practices. Computer Adaptive Testing (CAT) shows promising prospects as replacement for conventional paper-and-pencil tests. In a DOD study at the Navy Personnel Research and Development Center, San Diego, researchers are working to develop the capability of employing computers in testing. The next chapter(s) will describe the progress of that research and propose some strategies for implementing CAT into the DOD recruiting system.

V. COMPUTERIZED ADAPTIVE TESTING

The current turmoil over intelligence testing highlights a need for responsiveness; flexibility is required for the future. Critics are calling for ways to measure mental ability that will be fairer, more precise, and more relevant to real life than those now in use. Additionally, legislators are directing greater disclosure to consumers. Not only must the uses, methodologies, score-analysis techniques, and other characteristics of the tests be revealed, but the questions and correct answers must be provided in many cases as well. New York and California lawmakers enacted legislation which seriously affected the testing industry in those states [Weiss 1979] and, following that lead, Congressmen introduced the "Truth in Testing Act of 1979" (see Appendix A) and the "Educational Testing Act of 1979" (see Appendix B).

Spokesmen for the testing industry maintain that required disclosure of test contents greatly increases the demand on resources and thus also the cost of test development. As test items are developed, researched, administered, and disclosed, more new items must be quickly developed in order to offset the "compromise" of items. This argument has been countered by charges that companies that specialize in preparation courses and study guides for standardized tests routinely have access to test questions anyway. They allegedly have employees take the examinations and copy or memorize items

on the test for subsequent reproduction. In this manner, tests could be entirely reconstructed [Nairn and Associates 1980]. Regardless of how tests become public, the ability to develop unique items which meet the needs of consumers must be maximized in the future.

There is also a growing pressure to increase the number of variables which can be measured. For example, "intelligence as a measureable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test" [Boring 1923]. The current challenge seems to be to go beyond Boring's definition. Critics argue that tests today do not measure creativity, experience, or idealism, variables that, some say, should greatly enhance our ability to predict performance in a variety of educational programs and occupations [Weiss 1979]. Among psychologists, there is a "brave new world of intelligence research" evolving [Psychology Today 1979].

The computer offers the promise of capacity and flexibility which the future of mental measurement requires. One author predicts a variety of truly remarkable ways to measure intelligence in the future [Rice 1979, p. 27].

Future test takers, for example, may be:

- listening to clicks in earphones while electrodes taped to their temples send brain responses to be analyzed by a computer;
- held, as infants, by their parents while they watch toy cars roll down a ramp and -- sometimes -- knock over toy dolls;

- tested on mental abilities that may be influenced by watching television regularly;

- describing whether or not they prepare their own lunch, and relating how many pupils in their class they know by name;

- deciding at age three and a half how they would respond if they were in a game with three children and only two wanted to play.

New ideas which employ the computer are being developed and tested. The concept of "evoked potential" is one such measure. The basic technique employs an electroencephalograph (EEG) to measure the brain's response to a variety of stimuli such as sounds or flashes of light. Recent advances in computer technology have made it possible for scientists to sort out the minute changes that occur when a series of repeated stimuli suddenly stops or changes pattern. A series of alternating loud and soft clicks of equal intensity and duration which is suddenly changed or stopped would elicit some response from the brain. The brain disregards a regular pattern as it loses its novelty; however, the sudden change recaptures the brain's attention. The brain's response, or its evoked potential, gives a distinct measure which can be statistically compared to the response and IQ data of normal subjects [Rice 1979].

The Brain Research Laboratories at New York University Medical Center recently developed a sophisticated "Quantitative Electrophysiological Battery," which includes not only EEG readings but also 30 other measures of the brain's electrical activity as well. The computerized brain diagnoses might

someday be used for testing the intelligence potential of school children. It could be useful for those too young to take written exams or those whose command of English or late verbal development hinders their performance on written tests [Rice 1979].

Rice also describes a third type of research which uses heart and other muscle responses to measure mental processes in infants. Using electrocardiograph readings and the observations of two hidden assistants, Kearsely, cited by Rice, measures the cardiac response and behavior patterns indicating a child's reaction to a series of events. The speed with which the child responds to changes in the routine of events, Kearsely believes, provides a relatively uncontaminated measure of mental ability.

The U.S. Navy has developed a computerized Graphic Information Processing (GRIP) battery of tests in which trainees who are destined for electronic communications assignments can demonstrate specific abilities such as the visual processing of information or the tracking of a target on a radar screen [Rice 1979].

Thus we observe that the computer is a vital part of research and practical application. Now we want to describe a specific new research effort which has revolutionary implications for aptitude testing, selection, and placement in the armed forces -- computerized adaptive testing (CAT).

In this chapter, we will briefly describe conventional test design, define adaptive testing, and report on the

status of research into CAT currently undertaken by the armed forces.

A. CONVENTIONAL TESTING

ASVAB is a test that consists of multiple-choice items that may be considered to be given to all individuals in a group, even though the ASVAB appears in several forms. The test items are aimed at the average prospective recruit or high school student in the population of test takers, those individuals who are nearing or have reached armed-service eligibility age. The question has been raised whether these test items are appropriate for individuals who deviate significantly from the average [Weiss and Betz 1973].

Essentially, there are two extremes in test construction which relate to item difficulty. In conventional test construction, a test designer selects items from a pool of avaialable items that are known to measure a given variable, such as word knowledge or arithmetic reasoning ability. Within the pool, items vary in difficulty so that the test designer must decide what configuration of items best suits the purpose of the test [McBride 1980b]. If all test items selected are very similar in difficulty, the test is said to be a "peaked" test. In the extreme case of peakedness, an intelligence test would have all items of the same level of difficulty [Nunnally 1967]. Such a test would discriminate very effectively over a narrow range of the variable, but would

discriminate poorly outside that range. A peaked test may be used to make fine discriminations in the vicinity of a cutting point; in the case of recruiting for the armed services, results could be used to determine whether to select an individual or not [McBride 1980b].

At the other extreme, tests may be constructed with items that are most unlike in levels of difficulty. Such a test, called a "uniform" test, will discriminate over a wide range of the variable being measured. However, the level of precision will be substantially below that of the peaked test at the latter's best point. The uniform test best serves to obtain information which may be used to assist in deciding on placement in jobs that require varying amounts of ability [McBride 1980b].

The dilemma in selecting items to construct a conventional test is to choose between high precision over a narrow range (peaked test) or moderate precision over a wide range (uniform test). It is usually impractical to design a test of sufficient length to do both. Either the test must be extremely long or the item difficulty must be tailored to each examinee's level [McBride 1980b].

Stanley [1971] asserts that the effective length of any test is considerably less than the total number of items for any given test taker. He further states that it is wasteful of time and money to administer all items to all examinees. Adaptive testing, a new form of testing to be described below, offers some flexibility here without loss of reliability.

Weiss and Betz [1973] have identified a number of problems with paper-and-pencil conventional tests in general. They are listed and described briefly:

Individual Tests (One examiner and one examinee)

- There is evidence that different examiners score items on individual tests in different ways, even though they are following the same instructions.
- The degree of rapport between tester and testee can influence the results of individual ability tests.

Group Testing

- Administrator variables may influence test scores by inadvertently arousing anxiety in the test taker.
- Different types of answer sheets may have effects on test results.
- The selection and sequencing of test items can influence test scores, both for the group as a whole and for certain individuals in a group.
- Timing and time limits may affect the scores by rewarding the faster individual who has a tendency to guess and penalizing the slower more accurate individual. They may also penalize the person who tends to become anxious, and time limits can contribute to undesirable failure stress.
- We have mentioned the use of standardized test items above. When items are too difficult for a given examinee, the possibility of chance success through guessing on multiple-choice tests also contributes

error differentially. Guessing reduces the reliability and validity of measurement for all subjects, but the increase in error is particularly pronounced for low-ability subjects.

Space does not permit a more detailed discussion of the problems outlined by Weiss and Betz above; however, their brief mention does provide the basis for discussion of the concept of adaptive testing, which seeks to overcome some of these problems.

B. ADAPTIVE TESTING

In adaptive testing (also referred to as sequential, branched, individualized, tailored, programmed, response-contingent), the test administrator attempts to select test items in such a way as to adjust the level of difficulty to the ability demonstrated by the examinee during testing. Some steps which may occur in adaptive testing are summarized here:

1. Examiner makes judgment of testee's ability in order to determine at which ability level to begin testing.
2. Order of subsequent item presentation depends upon examinee's performance on previous items.
3. Extent of item presentation is controlled by basal and ceiling ages such that few items which are either too hard or too easy are presented.
4. Examiner provides encouragement and feedback to maintain examinee motivation.
5. No set time limits are imposed.

The Stanford-Binet [Terman and Merrill 1960] is considered to be one of the best representations of individualized adaptive tests because it is flexible enough to accommodate individual differences in ability and reaction to the testing process [Thorndike and Hagen 1977].

Individually administered adaptive tests are desirable in many respects because they afford flexibility and responsiveness; but they retain the qualities of subjectivity and susceptibility to administrator variables which render them unsatisfactory in terms of traditional psychometric criteria [Weiss and Betz 1973]. Additionally, conventional individual tests, such as the Wechsler Adult Intelligence Scale [Wechsler 1955], while providing individualized measurement, suffer from the lack of adaptive flexibility. Thus problems remain to be resolved which neither form of testing has yet managed. Research into the use of computerized adaptive testing is concerned in large part with solving these problems.

C. COMPUTERIZED ADAPTIVE TESTING

We have mentioned several areas in which computers function in ability measurement. Research is currently in progress at the Navy Personnel Research and Development Center in San Diego, California, to develop Computerized Adaptive Testing (CAT) for use in the selection and placement of military personnel.

In January, 1979, the Assistant Secretary of Defense (Manpower Reserve Affairs and Logistics) issued a

memorandum[1] for the Secretaries of the Army, Navy, and Air Force which discusses CAT. In part the memorandum says:

> Recent advances in the area of Computer Adaptive Testing (CAT) indicate that this is a promising approach with considerable potential for Defense selection and classification testing through the Armed Forces Examining and Entrance Stations (AFEES) system. Due to the Department of the Navy's R&D expertise in this area, Navy is designated lead service for the additional R&D required for the development and further evaluation of the feasibility of implementing CAT in the Department of Defense.

The memorandum served to stimulate more interest in CAT [McBride 1980a].

CAT is an attempt to employ automated, interactive terminals in the place of human examiners to administer adaptive tests. There are several significant benefits to be realized from CAT, of which four will be mentioned here. First, automated testing terminals should overcome the problem concerning how to conduct large-scale administration of adaptive tests. Traditional adaptive tests -- administered one-to-one -- are impractical on a large-scale basis; more than a million people are tested on the ASVAB each year. Presentation of test items and the recording of results can be done using computers linked with interactive devices such as teletype, cathode ray tube (CRT), or specially designed terminals, thus obviating the need for one-to-one human interplay. Second, with CAT it is possible to minimize

---

[1]Memorandum from the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics), Subject: "Computer Adaptive Testing," dated 5 January 1979.

the effects of the administrator variables which affect test scores [Weiss 1973]. (The question of the acceptability of machines to examinees has been raised and preliminary research at the U.S. Civil Service Commission indicates an overwhelmingly positive response to CAT [Schmidt, Urry and Gugal 1978]). Third, there should be a dramatic reduction of testing time via the tailoring of tests to the individual. Fourth, CAT should provide vast improvement in the cost, lead time, etc., of developing new test material.

1. CAT Procedure

The procedure for testing using CAT presumably will begin much as the ASVAB does except that interactive devices replace paper and pencil. Following some preliminary instructions from a "proctor," the examinees will begin testing by answering questions flashed on item-display devices. The essential components of the testing system are as follows [McBride 1980b]:

- Stimulus/display device
- Response device
- Item storage medium
- Internal processing
- Response processing capability
- Item selection capability
- Test scoring capability
- Data recording capability

Test items which are presented to each individual are initially selected on the basis of prior estimates of the

ability level. Such estimates may be determined in a variety of ways. For example, prior test scores may be useful for estimating an entry or "starting" ability-level point or perhaps high school grade-point averages may be employed; and there is the possibility of using a short "locator" test to approximate a beginning level or ability. Following the work of Binet, the idea is to enter the testing sequence at a level near the "basal" ability of the testee [Weiss 1973] -- that is the point below which further ability testing would yield no more information since all items would be answered correctly.

The testing sequence proceeds from the entry point through a series of questions. As each question is answered correctly or incorrectly, the next question may be of greater or less difficulty, respectively. Each variable may be measured in this manner until a ceiling level of ability -- where the testee answers all questions incorrectly -- is attained. The totality of test items between any testee's basal and ceiling ages can provide accurate measurement for that individual; for someone with different basal and/or ceiling levels, a different set of items will provide maximum information on his or her ability level [Weiss 1973]. Using the advantage of peaked test items, the process very efficiently -- in a few items -- can estimate the value of each variable. But, as we see in the discussion of research issues below, it is not clear whether this method, which

resembles the Binet scoring strategy, will necessarily be the scoring strategy for CAT.

2. Research at San Diego

Four CRT computer terminals are providing the interactive capability to communicate with a computer located at the University of Minnesota which provides the time-shared support for CAT research. These terminals, located at the U.S. Marine Corps Recruit Depot, San Diego, are being used to test the feasibility of CAT by testing Marine recruits. As described above, the recruits receive some brief preliminary instructions, identifying data is entered in the system, and the young recruits begin taking the tests by responding to instructions and questions displayed on the CRT's. The data gathered from these tests will be used to evaluate the performance of CAT in comparison with conventional tests of the subjects and in the resolution of other research issues.

D. RESEARCH ISSUES

There are a number of issues relevant to the employment of CAT on a routine basis. Earlier adaptive-testing research showed that traditional test theory was inadequate for the construction and scoring of CAT. McBride [1980b] identified a number of research issues which must be resolved. They are summarized below:

- Item Response Models. Within item response theory, several competing response models for dichotomously scored items have been proposed for adaptive testing.

These models differ in mathematical form and in the number of parameters needed to account for item response behavior. The issue which requires resolution is to choose the "best" model based upon its appropriateness, robustness under violations of relevant assumptions, and the difficulty and expense of implementing it.

- Strategies for Constructing Adaptive Tests. Since adaptive tests require that test items be selected sequentially, the best methods for selection, called "strategies," must be chosen. Research is needed to provide the basis for the best selection among competing strategies.
- Test Length. Conventional group paper-and-pencil tests are of fixed length. CAT affords the opportunity for varied or fixed-length tests. Research issues here concern the relative merits of fixed versus variable length -- whether variable-length tests have psychometric and practical merit.
- Test Entry Level. Already mentioned above, test entry level concerns the difficulty level of the first item(s) the examinee must answer. In theory, it seems beneficial to employ differential entry levels associated with each individual's respective abilities; however, more research is needed to affirm this.
- Scoring CAT. Adaptive tests present unique challenges in scoring. Conventional tests are primarily scored

by weighting and summing dichotomous -- 1 for correct answer, 0 for incorrect, for example -- item scores. Adaptive tests which may be of variable or fixed length, and which will vary in levels of difficulty among testees, may require significantly different scoring methods. The problem is two-fold: how can we score adaptive tests, and do methods exist to convert adaptive scores to conventional measures?

- The Testing Medium. CAT, by definition, will be administered by machine; we stated above that individual administration of adaptive tests by human examiners was impractical. While it may be feasible to administer adaptive tests by paper and pencil, the task of constructing such tests is formidable because of the requirement for sequential item selection.

Computer terminals and computers are expensive. The need for testing capability is defined by the summing of the number of testees which AFEES and mobile examining teams (MET) test each year. The potential expense of full computer support for each of the 67 AFEES's and 750 MET's presents a major hurdle. The issues are: (1) do we need full computer support for each testing site? (2) can less sophisticated devices be constructed for test administration? (3) what alternative devices/systems may be used for automated adaptive testing?

Note: McBride has developed a short adaptive test of mathematics skills which uses a hand-held programmable calculator. While the test is somewhat less sophisticated than what may be required in full-range adaptive tests, it demonstrates the idea of alternatives to full-scale computer adaptive tests. Contracted researchers are working on the device-development problem. Microprocessors or mini-computers may provide the answer.

- Item Pool Development. Since items are selected sequentially from a pool of items, the number of items must be substantially larger than that of a conventional test. Research must define the necessary characteristics of item pools.
- Advances in Measurement Methodology. CAT research may lead to further advances in ability measurement.

McBride [1980] presents a review of a variety of research efforts aimed at resolving the issues above and appears very optimistic about future use of adaptive testing in the military.

> The future psychometric and practical potential of adaptive testing makes it worthy of research and development in the military manpower setting, with the goal of eventual implementation of an automated system for test administration and scoring, and personnel selection, classification, and job-choice counseling [p. 4].

The goal for implementation of CAT is calendar year 1983 [McBride 1980c]. Because of the research issues still to be resolved above and the turmoil concerning testing in general, it seems appropriate to conclude the review of mental testing in the military with some statements concerning how the innovative computerized adaptive testing might be implemented. The next chapter will discuss organizational strategy associated with this implementation.

VI. SOME IMPLEMENTATION CONSIDERATIONS

Even in the face of the current turmoil concerning testing, it seems unlikely that decision-makers are going to readily relinquish the information-gathering processes which have enabled them to make "better" personnel decisions in the past. We are talking about how people differ and how we make decisions concerning who should get a job, who should be promoted, or who should be trained for technical or increased leadership roles. These are important decisions since people do differ in abilities, and wrong choices can have significant social and economic impact on individual as well as "organizational" lives. Wise decisions where people are concerned demand informed assessment of their individuality and abilities. Computerized adaptive testing is concerned with aiding the assessment of individual differences, so that selection, training, and placement decision-making may be improved in the military services.

The issues facing mental testing are real issues, and while there are some significant "problems" with psychometrics, we still must make decisions about people. To avoid the use of information which aids in the decision-making process would be to leave the futures of our institutions to chance. According to Dunnette [1966], "individual diagnosis must always be the crucial first step, undergirding and directing all subsequent personnel decisions."

The more relevant and accurate information we get about people's abilities, the better the decisions we can make. The periodic review and revision of the ASVAB and CAT research are efforts directed toward improving psychometric assessment capabilities. Measurement instruments and procedures provide an important set of tools for improving the information available for decision-making.

As stated in the previous chapter, the target date for implementation of CAT is 1983. In this chapter, we will list the major "players" in the implementation process and propose some considerations for managing the change from ASVAB to CAT.

A. MAJOR "PLAYERS"

The Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics (ASDM,RA&L), as stated in the previous chapter, has created the current impetus behind the research into CAT at San Diego. We noted above one role of that office, that of representative of DOD to Congress, in the testimony on manpower-quality issues. It is anticipated that the ASD(M,RA&L) will continue to take a vital interest in CAT particularly as it moves closer to implementation. Given that there is a high probability of continued debate over testing as the pending "truth in testing" bills are considered, ASD may play a major role in defending military mental testing.

Congress' role in CAT is not clearly defined at this time. Legislation which directs military qualification testing is already enacted; therefore, the services are able to proceed with development and administration of tests. However, if testing issues become the subject of increased public concern during the consideration of pending truth-in-testing legislation, the lawmakers may take a more active interest. It does not seem likely that CAT will be implemented without congressional scrutiny.

The Administrative chain of command which is overseeing CAT research consists of the following [McBride 1980]:

Deputy Assistant Secretary of Defense
(Military Personnel Policy)

Deputy Assistant Secretary of the Navy (Manpower)

Headquarters, United States Marine Corps, Chairman, Joint Services Committee, Computerized Adaptive Testing Interservice Coordinating Committee

Navy Personnel Research and Development Center,
San Diego, California

The Military Enlistment Processing Command (MEPCOM) is the potential user of CAT and a member of the coordinating committee.

The services are assigned specific responsibilities as follows [ASD memo]:

Department of the Navy:

- Chair and participate in a CAT inter-service coordinating committee for determining the feasibility and cost advantages of utilizing CAT in the Department of Defense. The committee is composed of representatives from the Services, Human Resources Research laboratories, the Military Enlistment Processing Command, appropriate Service policy personnel and the Civil Service Commission.

- Prime responsibility for development of psychometric methodology.

- Provide a "test bed" for testing of items and procedures during the program development phase.

Department of the Army:

- Participate as a member of the CAT coordinating committee.

- Have the prime responsibility for procurement and/or lease of the delivery system for CAT, if proven feasible and cost effective.

- Have the prime responsibility for possible implementation of CAT through the AFEES system.

Department of the Air Force:

- Participate as a member of the CAT coordinating committee.

- Have the prime responsibility for the development of item banks for the CAT Armed Services Vocational Aptitude Battery (ASVAB).

As noted above, the Marine Corps was designated lead Service within the Department of the Navy to execute the responsibilities identified in the ASD(MRA&L) memorandum [SECNAV MEMO]. Thus the players are each assigned roles; these roles are not unlike those of previous evolutions of military mental testing.

B. A MODEL OF CHANGE

The key players in the change from paper-and-pencil testing to computers will need a framework on which to base decisions about the implementation process. The following brief discussion will outline a model for change which may prove useful.

In Chapter IV we listed some issues/problems which confront testing; they are true for military mental measurement as well. One question which comes to mind is this: "If invasion of privacy is an issue in paper-and-pencil testing, how much more of an issue will it become in computerized testing?" Additionally we must keep in mind the fairness, test bias, and other issues which evolve.

Beckhard and Harris [1977] identify three stages which occur in introducing change into an organization, (1) the present state of affairs (prechange state), (2) the immediate goals of the organization, (3) and the desired final stage (postchange state). What the organization does during the change process must be managed; this is called the "transition state" [p. 5].

The transition state is considered to be dynamic and thus is a state of affairs in itself. For CAT research management, this implies that the functions associated with testing, selection, placement, and training of military personnel while we await CAT must continue. Management will need to deal with the ongoing affairs of DOD. The "quality issue," for example, must still be addressed, and efforts to improve the ASVAB paper-and-pencil forms must continue. The problem, as Beckhard and Harris state, is that we frequently begin to plan by detailing how people will behave and be organized once the change is complete but ignore the management exigencies of the transition state; we assume away the need to guide the organization through this phase.

In the case of the computerized adaptive testing (CAT) research, we have an inter-service committee consisting of representatives of constituencies -- that is, organizations which have a vested interest in military testing -- to manage the research and development effort. This organization should prove beneficial by insuring continuity during transition. Some of the items this committee must consider in order to manage the dynamic transition from paper and pencil to CAT, assuming it is determined to be feasible and economical, are the following:

- To what extent will society resist a break with traditional paper-and-pencil testing? Will test takers complain strongly about invasion of privacy? How will this be managed?
- What will be required to convince test users of the value of CAT? Of its accuracy as a measure?
- What will be required to gain the support of all the constituent organizations?
- What new technical skills and knowledge will be required for users of CAT? When should training begin? Who should be trained?

Detailed analysis of the questions raised above are beyond the scope of this paper. The questions are presented as a summary of those issues facing the major decision-makers in the military testing milieu. The significant burden on management is to manage the status quo and at the same time

provide direction during the dynamic evolution which is moving toward what may be a revolution in ability assessment.

Computerized Adaptive Testing possesses the potential for future growth in conjunction with other systems; electronic innovations which have measurement applications are being discovered or invented almost daily. The computer's increased capacity for processing and storing data and its ability to interface with other systems present potentially rewarding and challenging prospects for the field of psychometrics in the future; military testing has the opportunity to continue to lead.

The opportunity exists for students at the Naval Postgraduate School to continue the study of Computerized Adaptive Testing (and related areas of psychometrics) as this innovative idea is being researched and developed. The research issues discussed above will need to be resolved which means that meaningful and contributory thesis topics must be addressed.

In preparation for innovation in psychometric theory, and in order to enable military managers to improve their manpower selection and placement skills, more emphasis should be placed on psychometrics in the management and human resources curricula at the Naval Postgraduate School. Consider the question: "Can it be that one of the principal contributing factors to the turmoil over intelligence testing is the inept use of test results caused by ignorance of the meaning of test scores?" The answer to this question is probably yes. What is needed

at the Naval Postgraduate School and other institutions that prepare managers for leadership roles is additional education in psychometric theory and selection and placement. The recommended goal is that which will enhance the ability assessment skills and foster additional research in testing theory. There exists an opportunity to turn the tide of distrust and confusion over intelligence testing which seems to grow from a lack of understanding by investing knowledge and technical skills in the students of management who must deal with tomorrow's technology.

APPENDIX A

96th Congress
1st Session

H.R. 3564

To require all educational admissions testing conducted through interstate commerce, and all occupational admissions testing (which affects commerce) to be conducted with sufficient notice of test subject matter and test results, and for other purposes.

IN THE HOUSE OF REPRESENTATIVES

April 10, 1979

Mr. Gibbons introduced the following bill; which was referred to the Committee on Education and Labor

---

A BILL

To require all educational admissions testing conducted through interstate commerce, and all occupational admissions testing (which affects commerce) to be conducted with sufficient notice of test subject matter and test results, and for other purposes.

Be it enacted by the House of Representatives of the United States of America in Congress assembled, that this Act may be cited as the "Truth in Testing Act of 1979."

Sec. 2. As used in this Act --

(1) the term "educational admissions test" means any test of aptitude or knowledge which --

(A) is administered to individuals in two or more States.

(B) affects or is conducted or distributed through any medium of interstate commerce, and

(C) is used as part or all of the basis for admitting or denying admission to an individual to any institution of higher education;

(2) the term "occupational admissions test" means any test which is used as part or all of the basis for admitting or denying admission to an individual to any occupation in or affecting interstate commerce;

(3) The term "test" includes any aptitude or achievement examination, whether written or oral, and includes any objective multiple choice, machine scored, essay, practical, performance, or demonstration examination;

(4) the term "test score" means the numerical value given to the test subject's performance on any test;

(5) the term "person" includes individuals, corporations, companies, associations, firms, partnerships, societies, joint stock companies, and agencies and instrumentalities of States and local governments; and

(6) the term "institution of higher education" has the meaning set forth in section 1201(a) of the Higher Education Act of 1965 (20 U.S.C. 1141(a)).

Sec. 3. The Congress hereby finds and declares that --

(1) testing of scholastic aptitudes and achievements has become a principal factor in the admission of individuals to public, as well as to private, institutions of higher education and that therefore equal opportunity under the law requires that testing be conducted in a manner which will ensure the equal rights and fair treatment of such individuals;

(2) testing of skills for entry into an occupation, whether of a professional, craft, or trade nature, is a critical factor governing the free flow of individual skills in interstate commerce and seriously affects the Nation's capability for economic growth; and

(3) the rights of individuals and the national interests can be protected without adversely affecting the proprietary interest of any entity administering tests by simple requirements governing proper prior notice to individuals of the subject matter to be tested and proper subsequent notice of test results and their uses.

Sec. 4. It is the purpose of this Act to prohibit the conducting of educational and occupational admissions tests unless such tests are administered in a manner to protect the

rights of the individuals tested and to grant a Federal cause of action to any individual adversely affected by the administration of any such test in violation of this Act.

Sec. 5. It is unlawful for any person to administer any educational or occupational admissions test to any individual unless such test is administered in accordance with the requirements of section 6 of this Act.

Sec. 6(a) Each applicant to take any educational or occupational admissions test shall be provided with a written notice which shall contain --

(1) a detailed description of the area of knowledge or the type of aptitude that the test attempts to analyze;

(2) in the case of a test of knowledge, a detailed description of the subjects to be tested;

(3) the margin of error or the extent of reliability of the test, determined on the basis of experimental uses of the test and, where available, actual usage;

(4) the manner in which the test results will be distributed by the testing entity to the applicant and to other persons; and

(5) a statement of the applicant's rights under subsection (b) of this section to obtain test results and related facts.

(b) Each individual who takes any educational or occupational test shall, at the request of the test subject, promptly upon completion of scoring of such test, be notified of --

(1) the individual's specific performance in each of the subject or aptitude areas tested;

(2) how that specific performance ranked in relation to the other individuals and how the individual ranked on total test performance;

(3) the score required to pass the test for admission to such occupation or the score which is generally required for admission to institutions of higher education;

(4) any further information which may be obtained by the individual on request.

(c) No educational or occupational admissions test which tests knowledge or achievement (rather than aptitude) shall be graded (for purposes of determining the score required to

pass the test for admission) on the basis of the relative distribution of scores of other test subjects.

Sec. 7(a) Whenever any person has administered or there are reasonable grounds to believe that any person is about to administer any educational or occupational admissions test in violation of this Act, a civil action for preventive relief, including an application for a permanent or temporary injunction, restraining order, or other order, may be instituted by the individual or individuals aggrieved. Upon application by the complainant and in such circumstances as the court may deem just, the court may appoint an attorney for such complainant and may authorize the commencement of civil action without payment of fees, costs, or security.

(b) In any action commenced pursuant to this section, the court, in its discretion, may allow the prevailing party, other than the United States, a reasonable attorney's fee as part of the costs.

(c) The district courts of the United States shall have jurisdiction of proceedings instituted pursuant to this Act and shall exercise the same without regard to whether the aggrieved party shall have exhausted any administrative or other remedies that may be provided by law.

Sec. 8. This Act shall be effective with respect to any test administered on or after January 1, 1979.

APPENDIX B

96th Congress
1st Session

H.R. 4949

To require certain information be provided to individuals who take standardized educational admissions tests, and for other purposes.

IN THE HOUSE OF REPRESENTATIVES

July 24, 1979

Mr. Weiss (for himself, Mrs. Chisholm, and Mr. Miller of California) introduced the following bill; which was referred to the Committee on Education and Labor.

A BILL

To require certain information be provided to individuals who take standardized educational admissions tests, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

Short Title

Section 1. This Act may be cited as the "Educational Testing Act of 1979."

Findings of Purpose

Sec. 2(a) The Congress of the United States finds

(1) education is fundamental to the development of individual citizens and the progress of the Nation as a whole;

(2) there is a continuous need to ensure equal access for all Americans to educational opportunities of high quality;

(3) standardized tests are a major factor in the admission and placement of students in postsecondary education and also play an important role in individuals' professional lives;

(4) there is increasing concern among citizens, educators, and public officials regarding the appropriate uses of standardized tests in the admissions decision of postsecondary education institutions;

(5) the rights of individuals and the public interest can be assured without endangering the proprietary rights of the testing agencies; and

(6) standardized tests are developed and administered without regard to State boundaries and are utilized on a national basis.

(b) It is the purpose of this Act --

(1) to ensure that test subjects and persons who use test results are fully aware of the characteristics, uses, and limitations of standardized tests in post-secondary education admissions;

(2) to make available to the public appropriate information regarding the procedures, development, and administration of standardized tests;

(3) to protect the public interest by promoting more knowledge about appropriate use of standardized test results and by promoting greater accuracy, validity, and reliability in the development, administration, and interpretation of standardized tests; and

(4) to encourage use of multiple criteria in the grant or denial of any significant educational benefit.

Information to Test Subjects and Postsecondary Educational Institutions

Sec. 3(a) Each test agency shall provide to any test subject in clear and easily understandable language, along with the registration form for a test, the following information:

(1) The purposes for which the test is constructed and is intended to be used.

(2) The subject matters included on such test and the knowledge and skills which the test purports to measure.

(3) Statements designed to provide information for interpreting the test results, including explanation of the test, and the correlation between test scores and future success in schools and, in the case of tests used for postbaccalaureate admissions, the standard error of

measurement and the correlation between test scores and success in the career for which admission is sought.

(4) Statements concerning the effects on and uses of test scores, including --

(A) if the test score is used by itself or with other information to predict future grade point average, the extent, expressed as a percentage, to which the use of this test score improves the accuracy of predicting future grade point average, over and above all other information used; and

(B) a comparison of the average score and percentiles of test subjects by major income groups; and

(C) the extent to which test preparation courses improve test subjects' scores on average, expressed as a percentage.

(5) A description of the form in which test scores will be reported, whether the raw test scores will be altered in any way before being reported to the test subject, and the manner, if any, the test agency will use the test score (in raw or transformed form) by itself or together with any other information about the test subject to predict in any way the subject's future academic performance for any postsecondary educational institution.

(6) A complete description of any promises or covenants that the test agency makes to the test subject with regard to accuracy of scoring, timely forwarding or score reporting, and privacy of information (including test scores and other information), relating to the test subjects.

(7) The property interests of the test subject in the test results, if any, the duration for which such results will be retained by the test agency, and policies regarding storage, disposal, and future use of test scores.

(8) The time period within which the test subject's score will be completed and mailed to the test subject and the time period within which such scores will be mailed to test score recipients designated by the test subject.

(9) A description of special services to accommodate handicapped test subjects.

(10) Notice of (A) the information which is available to the test subject under section 5(a)(2), (B) the rights of the test subject under section 6, and (C) the procedure for appeal or review of a test score by the test agency.

(b) Any institution which is a test score recipient shall be provided with the information required by subsection (a). The test agency shall provide such information with respect to any test prior to or coincident with the first reporting of a test score or scores for that test to a recipient institution.

(c) The test agency shall immediately notify the test subject and the institituions designated as test score recipients by the test subject if the test subject's score is delayed ten calendar days beyond the time period stated under subsection (a)(8) of this section.

Reports and Statistical Data and Other Information

Sec. 4(a)(1) In order to further the purposes of this Act, the following information shall be provided to the Commissioner by the test agency:

(A) Any study, evaluation, or statistical report pertaining to a test, which a test agency prepares or causes to be prepared, or for which it provides data. Nothing in this paragraph shall require submission of any reports or documents containing information identifiable with any individual test subject. Such information shall be deleted or obliterated prior to submission to the Commissioner.

(B) If one test agency develops or produces a test and another test agency sponsors or administers the same test, a copy of their contract for services shall be submitted to the Commissioner.

(2) All data, reports, or other documents submitted pursuant to this section will be considered to be records for purposes of section 552(a)(3) of title 5, United States Code.

(b) Within one year of the effective date of this Act, the Commissioners shall report to Congress concerning the relationship between the test scores of test subjects and income, race, sex, ethnic, and handicapped status. Such report shall include an evaluation of available data concerning the relationship between test scores and the completion of test preparation courses.

Promoting a Better Understanding of Tests

Sec. 5(a) In order to promote a better understanding of standardized tests and stimulate independent research on such tests, each test agency --

(1) shall, within thirty days after the results of any standardized test are released, file or cause to be filed in the office of the Commissioner --

(A) a copy of all test questions used in calculating the test subject's raw score;

(B) the corresponding acceptable answers to those questions; and

(C) all rules for transferring raw scores into those scores reported to the test subject and postsecondary educational institutions together with an explanation of such rules; and

(2) shall, after the test has been filed with the Commissioner and upon request of the test subject, send the test subject --

(A) a copy of the test questions used in determining the subject's raw score;

(B) the test subject's individual answer sheet together with a copy of the correct answer sheet to the same test with questions counting toward the test subject's raw score so marked; and

(C) a statement of the raw score used to calculate the scores already sent to the test subject if such request has been made within ninety days of the release of the test score to the test subject.

The test agency may charge a nominal fee for sending out such information requested under paragraph (2) not to exceed the marginal cost of providing the information.

(b) This section shall not apply to any standardized test for which it can be anticipated, on the basis of past experience (as reported under section 7(2) of this Act), will be administered to fewer than five thousand test subjects nationally over a testing year.

(c) Documents submitted to the Commissioner pursuant to this section shall be considered to be records for purposes of section 552(a)(3) of title 5, United States Code.

Privacy of Test Scores

Sec. 6. The score of any test subject, or any altered or transferred version of the score identifiable with any test subject, shall not be released or disclosed by the test agency to any person, organization, association, corporation,

postsecondary educational institution, or governmental agency or subdivision unless specifically authorized by the test subject as a score recipient. A test agency may, however, release all previous scores received by a test subject to any currently designated test score recipient. This section shall not be construed to prohibit release of scores and other information in a form which does not identify the test subject for purposes of research leading to studies and reports primarily concerning the tests themselves.

Testing Costs and Fees to Students

Sec. 7. In order to ensure that tests are being offered at a reasonable cost to test subjects, each test agency shall report the following information to the Commissioner:

(1) Before March 31, 1981, or within ninety days after it first becomes a test agency, whichever is later, the test agency shall report the closing date of its testing year. Each test agency shall report any change in the closing date of its testing year within ninety days after the change is made.

(2) For each test program, within one hundred and twenty days after the close of the testing year, the test agency shall report --

(A) the total number of times the test was taken during the testing year;

(B) The number of test subjects who have taken the test once, who have taken it twice, and who have taken it more than twice during the testing year;

(C) the number of refunds given to individuals who have registered for, but did not take, the test;

(D) the number of test subjects for whom the test fee was waived or reduced;

(E) the total amount of fees received from the test subjects by the test agency for each test program for that test year;

(F) the total amount of revenue received from each test program; and

(G) the expenses to the test agency of the tests, including --

(i) expenses incurred by the test agency for each test program;

(ii) expenses incurred for test development by the test agency for each test program; and

(iii) all expenses which are fixed or can be regarded as overhead expenses and not associated with any test program or with test development;

(3) If a separate fee is charged test subjects for admissions data assembly services or score reporting services, within one hundred and twenty days after the close of the testing year, the test agency shall report --

(A) the number of individuals registering for each admissions data assembly service during the testing year;

(B) the number of individuals registering for each score reporting service during the testing year;

(C) the total amount of revenue received from the individuals by the test agency for each admissions data assembly service or score reporting service during the testing year; and

(D) the expenses to the test agency for each admissions data assembly service or score reporting service during the testing year.

Regulations and Enforcement

Sec.8(a) The Commissioner shall promulgate regulations to implement the provisions of this Act within one hundred and twenty days after the effective date of this Act. The failure of the Commissioner to promulgate regulations shall not prevent the provisions of this Act from taking effect.

(b) Any test agency that violates any clause of any provision of this Act shall be liable for a civil penalty not to exceed $2,000 for each violation.

(c) If any provision of this Act shall be declared unconstitutional, invalid, or inapplicable, the other provisions shall remain in effect.

Definitions

Sec. 9. For purposes of this Act --

(1) the term "admissions data assembly service" means any summary or report of grades, grade point averages, standardized test scores, or any combination of grades and test scores, of any applicant used by any postsecondary educational institution in its admissions process;

(2) the term "Commissioner" means the Commissioner of Education;

(3) the term "postsecondary educational educational institution" means any institution providing a course of study beyond the secondary school level and which uses standardized tests as a factor in its admissions process;

(4) the term "score reporting service" means the reporting of a test subject's standardized test score to a test score recipient by a testing agency.

(5) the term "standardized test" or "test" means --

(A) any test that is used, or is required, for the process of selection for admission to postsecondary educational institutions or their programs, or

(B) any test used for preliminary preparation for any test that is used, or is required, for the process of selection for admission to postsecondary educational institutions or their programs,

which affects or is conducted or distributed through any medium of interstate commerce, but such term does not include any test designed solely for nonadmission placement or credit-by-examination or any test developed and administered by an individual school or institution for its own purposes only;

(6) the term "test agency" means any person, organization, association, corporation, partnerhsip, or individual who develops, sponsors, or administers a standardized test;

(7) the term "test preparation course" means any curriculum, course of study, plan of instruction, or method of preparation given for a fee which is specifically designed or constructed to prepare a test subject for, or to improve a test subject's score on, a standardized test;

(8) the term "test program" means all the administrations of a test of the same name during a testing year;

(9) the term "test score" means the value given to the test subject's performance by the test agency on any test, whether reported in numerical, percentile, or any other form.

(10) the term "test score recipient" means any person, organization, association, corporation, postsecondary educational institution, or governmental agency or subdivision to which the test subject requests or designates that a test agency reports his or her score;

(11) the term "test subject" means an individual to whom a test is administered; and

(12) the term "testing year" means the twelve calendar months which the test agency considers either its operational cycle or its fiscal year.

Effective Date

Sec. 10. This Act shall take effect one hundred and eighty days after the date of its enactment.

LIST OF REFERENCES

References for Section II

Binet, Alfred and Simon, Theodore, The Development of Intelligence in Children, translated by Elizabeth S. Kite, ARNO Press, 1973.

Brubacher, John S., A History of the Problems of Education, McGraw-Hill, 1947.

Cattell, James McKeen, James McKeen Cattell: Man of Science, V. 1, Science Press, 1947.

DuBois, Philip H., A History of Psychological Testing, Allyn and Bacon, 1970.

Dunnette, Marvin D., Personnel Selection and Placement, Wadsworth Publishing Company, 1966.

Forrest, D. W., Francis Galton: The Life and Work of a Victorian Genius, Taplinger Publishing, 1974.

Heidbreder, Edna, Seven Psychologies, D. Appleton-Century, 1933.

Linden, Kathryn W. and Linden, James D., Modern Mental Measurement: A Historical Perspective, Houghton Mifflin Company, 1968.

Mayer, Frederick, A History of Educational Thought, Charles E. Merrill Books, 1960.

McGucken, W. J., The Jesuits and Education, Bruce Publishing Co., 1932.

Psychometric Methods Program, Department of Psychology, University of Minnesota, Research Report 73-3, The Stratified Adaptive Computerized Ability Test, by David J. Weiss, September, 1973.

Psychometric Methods Program, Department of Psychology, University of Minnesota Research Report 73-1, Ability Measurement: Conventional or Adaptive?, by David J. Weiss and Nancy E. Betz, February, 1973.

Spearman, Charles, "'General Intelligence,' Objectively Determined and Measured," in Readings in Human Intelligence. Butcher, H. J. and Lomax, E. E., Ed. Methuen and Co., Ltd., 1972.

Stern, William, "On the Mental Quotient," Translated by Guy Montrose Whipple in A Source Book in the History of Psychology, ed. by Richard J. Herrnstein and Edwin G. Boring, Harvard University Press, 1965.

Terman, Lewis M., The Measurement of Intelligence, Houghton Mifflin Company, 1916.

Tuddenham, Read D., "The Nature and Measurement of Intelligence," in Psychology in the Making, ed. by Leo Postman, Alfred A. Knopf, Inc., 1962.

White, Robert W., The Abnormal Personality, Ronald Press Company, 1964.

Yerkes, Robert M., "Psychology in Relation to the War," in American Psychology in Historical Perspective, ed. by Ernest R. Hilgard, American Psychological Association, Inc., 1978.

Yerkes, R. M., ed., "Psychological Examination of the Soldier," in The Harvey Lectures. Philadelphia and London: J. B. Lippincott, 1920, pp. 181-215.

References for Section III

Bingham, Walter V., "Psychological Services in the United States," Journal of Consulting Psychology, v. 5, p. 223, 1941.

Brandt, Hyman, "Development and Construction of an Armed Services Qualification Test: I. Rationale, Item Content and Construction," American Psychology, v. 4, p. 239, 1949.

Bray, Charles W., Psychology and Military Proficiency, Princeton University Press, 1948.

Cronbach, Lee J., "The Armed Services Vocational Aptitude Battery -- A Test Battery in Transition," Personnel and Guidance Journal, p. 232-37, January, 1979.

Davis, Robert A., "Testing in the Army and Navy," The Journal of Educational Psychology, v. 34, pp. 440-446, 1943.

Directorate of Testing, United States Military Enlistment Processing Command, Technical Research Report 77-4, Validity of the Armed Services Vocational Aptitude Battery (ASVAB) for Predicting Performance in Service Technical Training Schools, Fischl, M. A., and others, 1978.

Directorate of Testing, United States Military Enlistment Processing Command, Technical Research Note 77-3, Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7), Jensen, H. E. Massey, I. H., and Valentine, L. D., Jr., 1977.

Freeman, Frank A., Mental Tests: Their History Principles and Applications, Houghton Mifflin Company, 1926.

Goodenough, Florence L., Mental Testing: Its History Principles and Applications, Holt, Rinehart and Winston, 1961.

Kamin, Leon J., The Science and Politics of I.Q., Lawrence Erlbaum Associates, 1974.

Marks, Russell, "Providing for Individual Differences: A History of the Intelligence Testing Movement in North America," Interchange, v. 7, 1976-77, p. 3-16.

Pastore, Nicholas, "The Army Intelligence Tests and Walter Lippmann," Journal of the History of the Behavioral Sciences, v. 14, p. 316-27, 1978.

Reisman, J. M., The Development of Clinical Psychology, Appleton-Century-Crofts, 1966.

Samelson, Franz, "World War I Intelligence Testing and the Development of Psychology," Journal of the History of the Behavioral Sciences, v. 13, pp. 274-81, 1977.

Staff, Personnel Research Section, Classification and Replacement Branch, The Adjutant General's Office, "Testing as a Part of Military Classification," Science, v. 97, pp. 473-478, 28 May 1943.

Stuit, Dewey B., Personnel Research and Test Development in the Bureau of Naval Personnel, Princeton University Press, 1948.

U.S. Army Behavior and Systems Research Laboratory, Technical Research Report 1161, The Armed Services Vocational Aptitude Battery, Bayroff, A. G., and Fuchs, E. F., Unclas., 1970.

U.S. Army Research Institute for the Behavioral and Social Sciences, Research Report 1179, Effectiveness of Selection and Classification Testing, Maier, Milton H., and Fuchs, E. F., Unclas., 1973.

U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Paper 289, Development of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 2 and 3, Seeley, L. C., Fischl, M. A., and Hicks, J. M., 1978.

Wilkens, Walter L., "Psychiatric and Psychological Research in the Navy before World War II," Military Medicine, v. 137(6), pp. 228-231, 1972.

References for Section IV

"Beyond I.Q.: An Introduction," Psychology Today, p. 24, September, 1979.

Cleary, T. Anne, and Hilton, Thomas L., "An Investigation of Item Bias," Educational and Psychological Measurement, v. 28, pp. 61-75, 1968.

Directorate of Testing, United States Military Enlistment Processing Command, Technical Research Note 77-3, Armed Services Vocational Aptitude Battery Development of (ASVAB) Forms 5, 6, and 7, Jensen, H. E., Massey, I. H., and Valentine, L. D., Jr., December, 1977.

Dubois, Philip H., "Varieties of Psychological Test Homogeneity," American Psychologist, v. 25, p. 532, 1970.

Dunnette, Marvin D., Personnel Selection and Placement, Brooks/Cole, 1966.

Educational Testing Service (ETS), "Test Use and Validity: A Response to Charges in the Nader/Nairn Report on ETS," Princeton, New Jersey, February 1980.

Flaugher, Ronald L., "The Many Definitions of Test Bias," American Psychologist, v. 33, p. 671, July 1978.

Green, Bert F., "In Defense of Measurement," American Psychologist, v. 33, p. 664-70, July 1978.

Herrnstein, R. J., "In Defense of Intelligence Tests," Commentary, pp. 40-51, February 1980.

Huff, Sheila "Credentialing by Tests or by Degrees: Title VII of the Civil Rights Act and Griggs V. Duke Power Company," Harvard Educational Review, v. 44, p. 246, 1974.

Interview with Professor Ronald Weitzman of Naval Postgraduate School, Monterey, November 20, 1980.

Interview with Professor R. A. Weitzman, Naval Postgraduate School, Monterey, California, November 26, 1980.

Jensen, Arthur R., Bias in Mental Testing, Macmillan, 1980.

Kamin, Leon J., The Science and Politics of I.Q., Lawrence Erlbaum Associates, 1974.

Nairn, Allan, and Associates, "The Reign of ETS: The Corporation that Makes up Minds," The Ralph Nader Report on the Educational Testing Service, Washington, D.C., 1980.

Philpott, Tom, "Services Told to Lower Scores on AFQT's," Navy Times, p. 18, November 3, 1980.

Philpott, Tom, "Mental Category Rules Could be Minority Barrier," Navy Times, p. 16, October 6, 1980.

Project 100,000, "Characteristics and Performance of 'New Standards' Men," Office of Secretary of Defense, December 1969.

Rice, Berkeley, "Brave New World of Intelligence Testing," Psychology Today, p. 27+, September, 1979.

Slack, Warner V., and Porter, Douglas, "The Scholastic Aptitude Test: A Critical Appraisal," Harvard Educational Review, v. 50, pp. 154-175, May 1980.

The New York Times, "Volunteer Army Hurt by Drop in Test Scores and Cuts in Training," p. 36:4, April 6, 1980.

Thorndike, Robert L., and Hagen, Elizabeth P., Measurement and Evaluation in Psychology and Education, 4th ed., John Wiley and Sons, 1977.

U.S. Congress. House. Committee on Appropriations, Department of Defense Appropriations for 1981, Hearings before a subcommittee of the Committee on Appropriations, House of Representatives, 96th Congress, 2nd sess., 1980.

Weiss, Ted, "Statement on the Educational Testing Act," Congressional Record, E3862, Washington, D.C., July 25, 1979.

References for Section V

"Beyond I.Q.: An Introduction," Psychology Today, p. 25, September 1979.

Boring, Edwin G., "Intelligence as the Tests Test It," The New Republic, v. 34, p. 35-36, 1923.

McBride, James R.

a. Interview with James R. McBride, Naval Research and Development Center, San Diego, California, 26 February 1980.

b. McBride, James R., Adaptive Mental Testing: The State of the Art. To be published as a Technical Report of the U.S. Army Research Institute for the Behavioral and Social Sciences, 1980.

c. Interview with James R. McBride, Naval Research and Development Center, San Diego, California, 29 October 1980.

Nairn, Allan, and Associates, "The Reign of ETS: The Corporation that Makes Up Minds," The Ralph Nader Report on the Educational Testing Service, Washington, D.C., 1980.

Nunally, Jim C., Psychometric Theory, McGraw Hill, 1967.

Psychometric Methods Program, Department of Psychology, University of Minnesota Research Report 73-3, The Stratified Adaptive Computerized Ability Test, by David J. Weiss, 1973.

Psychometric Methods Program, Department of Psychology, University of Minnesota Research Report 73-1, Ability Measurement: Conventional or Adaptive?, by David J. Weiss and Nancy E. Betz, February, 1973.

Rice, Berkeley, "Brave New World of Intelligence Testing," Psychology Today, pp. 27-41, September, 1979.

Schmidt, Frank L., Urry, Vern W., and Gugel, John F., "Computer Assisted Tailored Testing: Examinee Reactions and Evaluations," Educational and Psychological Measurements, v. 38, pp. 265-273, 1978.

Stanley, J. C., "Reliability," in Educational Measurement, R. L. Throndike (Ed.), American Council on Education, Washington, D.C., 1971.

Terman, L. M., and Merrill, M. A., Stanford-Binet Intelligence Scale, Houghton Mifflin, 1960.

Thorndike, Robert L. and Hagen, Elizabeth P., Measurement and Evaluation in Psychology and Education, 4th ed., John Wiley and Sons, 1977.

Wechsler, C., Wechsler Adult Intelligence Scale, The Psychological Corporation, 1955.

Weiss, Ted, "Statement on the Educational Testing Act," Congressional Record, E 3862, Washington, D.C., July 25, 1979.

References for Section VI

Beckard, Richard, and Harris, Reuben, Organization Transition : Complex Change, Addison-Wesley, 1977.

Dunnette, Marvin D., Personnel Selection and Placement, Brooks/Cole Publishing Company, 1966.

Interview with James R. McBride, Navy Personnel Research and Development Center, San Diego, 29 October 1980.

Secretary of the Navy Memorandum: Subject: "Computer Adaptive Testing," dated 29 January 1979.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 54 Department of Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
4.	Professor R. A. Weitzman, Code 54 Wz Department of Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
5.	Assistant for Analysis, Evaluation (NMPC-6C) Human Resource Management & Personal Affairs Department Navy Military Personnel Command Washington, D.C. 20370	1
6.	Director, Human Resource Management Division (NMPC-62) Human Resource Management & Personal Affairs Department Navy Military Personnel Command Washington, D.C. 20370	1
7.	Director for HRM Plans and Policy (OP-150) Human Resource Management Division Deputy Chief of Naval Operations (Manpower, Personnel & Training) Washington, D.C. 20370	1
8.	Commanding Officer Human Resource Management School Naval Air Station Memphis Millington, Tennessee 38054	1
9.	Commanding Officer Human Resource Management Center London Box 23 FPO New York 09510	1

10. Commanding Officer — 1
Human Resource Management Center
5621-23 Tidewater Drive
Norfolk, Virginia 23509

11. Commanding Officer — 1
Human Resource Management Center
Pearl Harbor, Hawaii 96860

12. Commanding Officer — 1
Human Resource Management Center
Naval Training Center
San Diego, California 92133

13. Commanding Officer — 1
Human Resource Management Center
Commonwealth Building, Room 1144
1300 Wilson Blvd.
Arlington, Virginia 22209

14. Commanding Officer — 1
Naval Hospital Corps School
Great Lakes, Illinois 60088

15. Dr. James R. McBride — 1
Navy Personnel Research and
Development Center
Code 310
San Diego, California 92152

16. Commanding Officer — 1
Health Sciences Education and Training Command
National Naval Medical Center
Bethesda, Maryland 20014

17. Chief, Bureau of Medicine and Surgery — 1
Navy Department
Washington, D.C. 20372

18. LCDR Robert S. Kayler, USC, USN — 1
613 Blossom Drive
Rockville, Maryland 20857*0*

19. Commanding Officer — 1
Naval School of Health Services
National Naval Medical Center
Bethesda, Maryland 20014